# REPORT RESUMES

ED 003 855 64
EXPERIMENTAL RESEARCH ON EDUCATIONAL MEDIA.
BY- LUMSDAINE, A.A. ROSHAL S.M.
CALIFORNIA UNIV., LOS ANGELES
REPORT NUMBER NDEA-VIIB-057 PUB DATE 63
REPORT NUMBER NDEA-VIIB-057-A
EDRS PRICE MF-$0.99 HC-$27.12 678P.

DESCRIPTORS- BOOKS, TEXTBOOKS, *REFERENCE BOOKS, *EDUCATIONAL RESEARCH, *EDUCATIONAL RESOURCES, *MEDIA RESEARCH, *AUDIOVISUAL INSTRUCTION, TELEVISED INSTRUCTION, PROGRAMED INSTRUCTION, FILMS, RESEARCH METHODOLOGY, RESEARCH PROBLEMS, RESEARCH PROJECTS, RESEARCH SKILLS, RESEARCH TOOLS, LOS ANGELES, CALIFORNIA

A "GUIDEBOOK" WAS DEVELOPED AS A CONTRIBUTION TO THE IMPROVEMENT OF EXPERIMENTAL RESEARCH ON EDUCATIONAL MEDIA. ONE OF ITS MAIN AIMS WAS TO PROMOTE BETTER UNDERSTANDING, AMONG BOTH SPONSORS AND INVESTIGATORS, OF THE ART IN EDUCATIONAL-MEDIA EXPERIMENTATION. THE BOOK GUIDES THE READER IN TWO WAYS--(1) BY PROVIDING A "PRIMER" OF SOME FUNDAMENTAL CONSIDERATIONS USEFUL FOR RESEARCHERS AND SPONSORS TO KEEP IN MIND AND (2) BY GUIDING THE READER TO FURTHER INFORMATION ABOUT RELEVANT ASPECTS OF EDUCATIONAL MEDIA RESEARCH. (JC)

ED003855

B-057
B-057a

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

# EXPERIMENTAL RESEARCH ON EDUCATIONAL MEDIA

by

A. A. Lumsdaine

and

S. M. Roshal

Prepared under Contract #SAE 9063 with the Educational Media Brand, U. S. Office of Education, Department of Health, Education, and Welfare, under the provisions of the National defense Education Act, Title VII Part B.

Los Angeles: University of California, 1963

Copyright 1963
A. A. Lumsdaine

# PREFACE

The preparation of this "guidebook" was undertaken as a contribution to the improvement of experimental research on educational media. One of its main aims is to promote better understanding, among both sponsors and investigators, of the current state of the art in educational-media experimentation.

The term "guidebook" is used advisedly. The book does not purport to be a comprehensive textbook. It does not cover all aspects of experimental research on educational media, and it does not provide a complete "handbook" of all the pertinent research and statistical data available. However, it does seek to provide useful guidance to the reader, in two main ways: (1) by providing a sort of "primer" to present some fundamental considerations which seem useful for research workers and sponsors to keep in the focus of attention, and (2) by guiding the reader to further information about relevant aspects of educational-media research.

In order to reach as broad as possible a range of relevant audiences, much of the book's content and presentation were chosen so as to be appropriate for the neophyte in educational-media experimentation. However, the more experienced research worker was also kept in mind, and it is felt that emphasis on elementary considerations need not detract from the book's value to the more mature investigator. Indeed, it is the writer's experience that more experienced research workers are often helped by refocusing their attention on some elementary considerations which sometimes are lost sight of as one becomes pre-occupied with a mass of technical and administrative details in conducting extended research projects.

The preparation of the book was made possible through contract support from the Educational Media Branch of the United States Office of Education, Department of Health, Education, and Welfare, under the provisions of Title VII, Part B, of the National Defense Education Act. The undersigned was responsible, as senior author and principal investigator, for the conception and basic planning of the volume. Several temporally overlapping and related activities fed into its preparation. A foundation for a major portion of an initial draft was provided through the writing of my chapter, "Instruments and Media of Instruction," for the Handbook of Research on Teaching (N. Gage, Ed., 1963). (I had also presented a briefer version of this material in February, 1960 at a conference sponsored by the Office of Naval Research at the University of Pittsburgh: see Chapter 9 in Training Research and Education (R. Glaser, Ed., 1962). A second contributing activity was the review and editing of a number of reports of instructional-media research, based on an extended series of studies earlier conducted under my direction as part of an Air Force research and development program during the period 1950-57. The editing of these Air Force reports, now available through their publication by the National Academy of Sciences in the volume Student Response in Programmed Instruction (Lumsdaine, 1961 x), also helped to identify and illustrate methodological points with which the present volume deals.

Dr. Sol M. Roshal, who collaborated on the present version of the book, has been associated with me in varying capacities for a number of years; in particular, he was my associate and principal deputy for a considerable portion of the above-mentioned Air Force research program. Some of the common points of view on research policy and design which we share, and which are reflected in the present volume, were initially shaped or consolidated during that period of association. More recently and directly, Dr. Roshal contributed to the present version of the volume as consultant and collaborator during the latter part of 1961 and the first part of 1962. In addition to less specific contributions in re-outlining and helping to organize some of the volume's conceptual and rhetorical framework, he wrote the basic drafts of Chapters 5 and 6 of the present version, and drafted what is currently the concluding section of Chapter 3 as well as the introductory part of Chapter 8 and short sections in other chapters.

A substantial obligation is owed by the writer to Nancy R. Hamilton, who as my senior research assistant during the period 1961-62 had responsibility for many phases of the work in organizing references and draft material, supervising the preparation and checking of the manuscript, and performing a variety of related tasks which helped greatly to lighten the burden of writing. I should also like to acknowledge my less direct but very real indebtedness to several of my past and present colleagues, and to a number of my students at the University of California whose instruction in my classes has seldom been a one-way street. Without the feedback they have provided, the presentation of concepts herein would in many respects have attained less clarity than has been achieved thus far.

My greatest obligation stems from a long-standing association with Fred D. Sheffield of Yale University. The present treatment of statistical inference and other aspects of research design and analysis has profited greatly from discussions and correspondence with Professor Sheffield extending over a period of many years. This applies not only in respect to particular topics and points presented, only a few of which are identified by specific footnote references, but also more broadly in terms of a basic stress on clarity of logical analysis which I have been able to reflect much less than perfectly here. In addition to this more general, long-range contribution, Professor Sheffield has also read the first ten chapters of the present draft, and has contributed a wealth of most helpful criticisms and suggestions which will materially improve many portions of the revised draft as it is edited for publication.

In addition, I should like to thank several other individuals who have read the manuscript and who have made helpful suggestions and comments. These include Dr. Merlin C. Wittrock, Dr. William H. Allen, Dr. W. C. Meierhenry, Dr. Warren Seibert and Dr. Wendell Smith. I also wish to express my appreciation for permission given to quote from numerous sources which will be more fully listed in the notes and Appendix of the final draft. For assistance in looking up and checking references, as well as for a considerable share of the typing and proofreading, my particular thanks are due to Nina Van Zoeren, who worked closely with Mrs. Hamilton and with Darlene Lloyd in these matters. I should also like to acknowledge gratefully the extensive assistance given by Gwen Austin, Suzanne Guerin and Tommy Armand in typing and proofreading the manuscript and reproduction masters for the present version of the book.

Los Angeles                                        A. A. Lumsdaine
Fall   1962

# CONTENTS

Chapter

1 Purpose and Scope

2 Educational Media as Objects of Research

3 The Nature of Experiments

4 Illustrative Hypothetical Experiments

5 Hypothesis Testing

6 Design Steps in Controlling Alternate Hypotheses

7 The Grammar of Instructional-Media Experiments

8 Dependent Variables and Their Assessment

9 Procedural Patterns of Instructional Experiments

10 Experimental Structure and Sampling

11 Stratification-Replication as Related to the Precision and Generality of Data

12 Strategy and Tactics in Instructional-Media Experimentation

13 The Research Literature on Educational-Media Experiments and on Student Response as a Factor in Instruction

14 Some Experimental Studies of Stimulus and Organization Variables in Instructional-Media

15 Some Special Problems in Instructional-Media Research

Appendix

Bibliography

# Chapter 1

# Purpose and Scope

An overview for the prospective reader

You may find this "primer and guidebook" helpful if you are going to:

- plan, design or write a proposal for a scientific experiment on educational media (or an experimental assessment of the effects of a particular educational-media "instrument" or program); or
- conduct an experiment, analyze its results or write a research report on it (either a detailed technical report, or an abstract or summary); or
- read a report of research on educational media, either in order to evaluate or review it, or in order to learn what has been found out (as a basis for general enlightenment, or for practical decision about the design or use of educational-media program); or
- read and evaluate a research proposal to determine whether you might wish to sponsor, support or participate in the research proposed; or
- formulate policy about programs of research and development of educational media.

If your activities and interests place you in one or more of these categories, the authors feel you could look to this book, not as an encyclopedia of all wisdom, nor even as a comprehensive textbook, but as a guidebook, which should help even the most experienced researcher to:

> avoid some rather prevalent misconceptions misplaced emphases and pitfalls in the design, execution or interpretation of results in experimental research on educational media;
>
> be better able to distinguish essential requirements from peripheral desiderata in designing, evaluating or interpreting educational-media experiments; or
>
> gain an added respect both for the importance and the difficulty of conducting good experiments;
>
> derive some helpful specific suggestions, as well as some general sophistication, that may improve the precision or validity of an experiment you may support or perform, or lessen the cost or improve the "yield" per unit cost of the experiment; and
>
> both save you some trouble and worry that are needless and also _cause_ you some trouble and worry that are needful if your experiments are to be worthwhile.

To the extent that one or more of these objectives is even partially contributed to by your use of the book, its purposes will have been in some measure realized. On the other hand, it would be fatuous to hope that it can fully solve your problems or provide a cure-all for all the ills that beset experiments from conception to burial (i.e., to publication in a technical journal without further utilization). At a minimum, the book should _help_ (more than hinder) in the making of good decisions about research on educational media. But it cannot foresee all the problems that will arise, and it cannot guarantee to provide the wisest advice (or even infallibly correct facts) on every point it touches. Above all--it cannot guarantee that researchers who heed its precepts will necessarily produce fruitful research. Ultimately, good research depends on good ideas, and we have very few good ideas on how to get good ideas for research. However, we do have some thoughts on how to spot or avoid what seems to be some forms of relatively fruitless ideas (and even a few hopeful hints on where one might start looking for good ones). In a word, it is hoped that the book will help what might otherwise be good research to be sometimes a little better, and will cause what might otherwise be worthless or mediocre research to be done a little less often.

The book deals primarily with instructional media as objects of experimental research, and in particular with the kinds of experimental research which can result in improved prediction and control of the effects of instructional media in attaining specific outcomes. The terms "experimental research" as here used is roughly synonymous with the idea of valid measurement of the effects produced by instructional progress. This characterization will be elaborated and somewhat sharpened presently.

Needs for research. A first question to ask is why we should be interested in research on educational media. Principally, this book was written in the hope that it could contribute to better instruction through fostering of better instructional-media research. Without derogating the value of knowledge as such, we are thus considering that the main eventual goal of such research is to improve instruction. This we will take as an end in itself, without recourse to other, more ultimate goals to justify improving instruction.

We recognize that there are many ways to improve instruction. Research is not the only way; instruction can be improved without better research. Nevertheless, it is the authors' conviction that scientific knowledge based on research is the best resource for permanently improving instruction. Without defending this position, sufficient to say that the state of a science generally limits the amount of improvement in practical application, and we feel that instruction is no exception. Moreover, in the absence of a good methodology for evaluating the effects of instruction, there is no way to tell whether one is improving instruction when innovations are introduced. For this methodology, we also have to look to research.

Why experimental research is needed

However, the particular field of research with which this book is primarily concerned goes beyond mere techniques for assessing the improvement of any particular instance of instruction or innovation in educational media. Although measurement methodology is necessarily given considerable attention, the primary concern of the book is, as stated above, with the conduct of experimental research. Such research can lead to greater knowledge and ability to predict the ways in which instruction using educational media can be made most effective. Basically, this involves testing hypotheses about factors that govern media effects. This notion of hypotheses testing leads to further questions.

First, granted that we want to improve instruction, why do we need to make experimental tests of hypotheses in order to guide us in making such improvements? Why is the extensive intuition, experience and genius of gifted teachers and media practitioners not sufficiant?

It is obvious that much successful and doubtless excellent instruction through media is accomplished without contributions from research. (Even so, as just noted, we at least used research-based measurement methodologies to determine the effects that the instruction has had, if we wish to remove the assessment of excellence in instruction from the realm of mere opinion and impression.) It is also evident that, in the preparation of any given instructional-media program, experienced teachers have much to contribute in terms of suggestions and guidance concerning the presentation of materials which they have had long experience in teaching, and that educational-media experts will similarly have much to contribute with respect to techniques of presentation which they have developed.

One way to state the need for experimental research to supplement and go beyond the fruits of past experience of practitioners, however, becomes apparent as soon as several "experts" from the teaching or media fields are asked to make explicit how a particular lesson should be presented or taught or how a particular instructional outcome should be achieved. Disagreements immediately become apparent; the experts disagree. Not all can be right; and one suspects that in some instances all may be wrong. Thus, if experimental research can provide techniques for testing the implications of divergent or contradictory approaches or incompatible hypotheses about how to teach a subject-matter through the use of educational media, it can make an important contribution to resolving uncertainties, reconciling divergent opinions, and providing a better basis for predicting what technique will work best in a particular situation. In addition, suppositions or hypotheses about the most effective way to use media in teaching a subject often lead, when put to empirical test, to unexpected results. Thus, such research affords a source of new hypotheses and can suggest new techniques that have not previously been in the stock and trade of the practitioner.

## Why research methods need to be improved

A further question that needs to be made explicit in understanding the rationale for the present book is to ask why, granted the foregoing, there is a need to improve knowledge of the techniques of experimental research. There are really two aspects to this question.

One aspect is that the best techniques of research that have thus far been developed are still quite imperfect. There are many methodological problems with which we have only at best moderately satisfactory techniques for dealing. For numerous problems in the rationale and design of experiments, as well as in the measurement of instructional outcomes, the best conceptions and methods thus far developed are still relatively crude either as compared with the precision of thought and technique found in the physical sciences, or in comparison with the evident needs of research in education. This will be made clearer as we go on; it is also evident to any critical and sophisticated reader on an examination of the current research literature.

The second aspect is that, unfortunately, even the best techniques now known are far from universally employed in current educational-media research. Indeed, in many instances gross errors and misconceptions have entered into the design of much of what passes for experimentation on educational media. The authors believe that there are in turn at least two reasons for this fact (which can be amply documented by an inspection of the research literature). A first reason is that many individuals have become interested in doing research on educational media and have gone ahead to carry out studies with commendable enthusiasm and industry but, unfortunately, without a solid grounding in experimental philosophy and technique. A second reason is that there has thus far been no systematic treatment addressed directly to the problems of rationale, method and technique in the experimental study of instructional media.

For this latter reason, many pitfalls and difficulties which a few more experienced researchers have learned to identify and avoid are not widely known; the same mistakes therefore keep getting made over and over again. On the positive side, various useful techniques wheih have been developed have not become well known; these are to be found in various individual research reports that have not been brought together in a single methodologically oriented source. A further way of delineating the purpose of this book is to say that it seeks to remedy the hiatus thus created.

Need for such a book as this. One further question might be raised. In view of the existence of several excellent works on experimental design, measurement and statistical method, why don't these sources suffice? The authors have felt that they do not suffice for several reasons. In the first place, there are many special problems and possibilities encountered in research on educational media which are not dealt with by more general works on experimental design and measurement. Related to this is the fact that these more general works, even where applicable, generally use illustrative material from fields (e.g., agriculture or biology) other than that of instructional media; thus the applicability of the methods they present is not always clearly apparent. In addition, many of the techniques presented, while useful in other fields, or to advanced researchers, are either not particularly useful in the field of instructional media experimentation or are presented at a level too difficult for the neophyte in instructional-media research.

Furthermore, it is unfortunate but true that, in many cases, individuals with very considerable advanced knowledge of statistical techniques--including complex analysis of variance and co-variance--nevertheless either do not fully grasp, or tend to lose sight of, some of the basic fundamentals without which the application of complex techniques to experimental design and analysis may be unnecessary and even wholly fallacious. This is even true of otherwise very competent teachers of statistics courses, some of whom are more conversant with various elaborate statistical refinements than are the present authors, but who can nevertheless be led far astray--perhaps by their very versatility--so as to lose sight of the fundamental requirements of statistical logic as applied to experimental data.

Finally, the fields of inferential statistics and experimental design are in themselves in process of evolution (and not infrequently fraught with controversy). The inexperienced experimenter on educational media is hardly in a position to keep up to date with, or to assess or resolve differences among new experimental approaches. While the authors of this book by no means would claim omniscience, infallibility or even always adequate expertness in this respect (as noted in the preface), they do feel that extensive experience in the application of a diversity of methods to the special problems of instructional-media experimentation at least offers a basis for some helpful guidance to the reader.

## Difficulties of Experimental Research

Research is a slow process; there is no easy, royal road to knowledge. Experiments are often unsuccessful; even when apparently successful at the time, the results one thinks he has demonstrated may be controverted by a later experiment. We cannot expect a single experiment to result in an educational "breakthrough"; at most we may hope it to result in a solid, albeit limited contribution to knowledge

The progress of research generally can't be hurried by trying to take big bites or "giant steps." Trying to make giant strides by working on great, grandiose, broadly-defined problems is generally self-defeating. The big, large-sounding problems usually aren't really amenable to scientific experimentation. Almost always they have to be redefined, and smaller pieces of them broken off for intensive inquiry before one can really find out much. Indeed, large, grandiose kinds of instructional experiments (which sometimes take the form of "cross media" comparisons) often end with no useful results.

Progress in research is also limited by the tools available. In physical-science research, the point has often been made that a new piece of instrumentation may mean a breakthrough in experimentation and in turn permit an advance in theory. Similarly, in educational-media research we are hampered at any point, by the state of the art--by our ability, for example, to measure certain kinds of instructional outcomes in an efficient manner.

The existence of difficulties, shortcomings and pitfalls in the conduct and interpretation of experimental research is by no means peculiar to educational experimentation. Perhaps the difficulties more generally inherent in conducting satisfactory experiments are indicated by the fact that one of the two quotations with which Sir Ronald A. Fisher chose to preface his classic work, The Design of Experiments, (fourth edition, 1947) was the following from Robert Boyle's essay, "Concerning the unsuccessfulness of Experiments":

> "I am very sorry, Pyrophilus, that to the many (elsewhere enumerated) difficulties which you may meet with, and must therefore surmount, the serious and effectual prosecution of experimental philosophy I must add one discouragement more, which will perhaps as much surprise as dishearten you; and it is, that besides that you will find (as we elsewhere mentioned) many of the experiments published by authors, or related to you by the persons you converse with, false and unsuccessful (besides this, I say), you will meet with several observations and experiments which though communicated for true by candid authors or undistrusted eye-witnesses, or perhaps recommended by your own experience, may, upon further trial, disappoint your expectation, either not at all succeeding constantly, or at least varying much from what you expected."

We believe that it will quickly become apparent to the researcher in educational media that these words, written by a physical scientist in 1673, are as applicable (if not more so) to educational media-experiments as to experiments in biology, chemistry, or physics, and that even the best of available techniques makes them hardly less applicable to educational experimentation today than they were twenty or thirty--or three hundred--years ago.

## Limitations of methodological advances

Improved methodology is only a necessary condition for improved research--not a sufficient one. The success of research ultimately depends on creativity--one might even say, "genius". This applies particularly with respect to the formulation of problems of hypotheses for experimental test, but it also applies with respect to having a "green thumb" in implementing experimental variables, and in devising ingeniously sensitive ways to measure the effects of instruction. As Tukey (1961, pp. 119-120) has recently pointed out: "The nature of intelligent experimentation is a subject intuitively understood by many scientists, yet little written about. When one knows thus-and-so, what is the wise experimenter to do next? Such questions seem not to be covered anywhere." Tukey further quotes Ingle (1958) saying, "'I wish that I could say more on the positive side to aid the student who will search this and other books in vain to learn how to deal with other than cut-and-dried problems.'" Unfortunately, we do not know much about how to teach creativity and hence we can not say very much about how to get bright ideas for experiments, (although some poverty-stricken suggestions are attempted here and there: cf. Chapter XVI).

As Wallis and Roberts

> "The methods most important... [in hypothesis formulation] are primarily those of intuition, insight, imagination, and ingenuity. Very little can be said about them formally; perhaps they can be learned, but they cannot be taught. As someone has said, referring to an apocryphal story, many men noticed falling apples before Sir Isaac Newton, yet no interpretations of comparable interest were recorded by these earlier observers."

### Summary of purpose

A way of summarizing the major purpose of this guidebook is that, to the extent possible and within the range of techniques which have thus far been developed, we want to try to communicate those things that can be learned, from the experience of others in doing experiments, which will enable future experimenters to minimize the errors that are conducive to unsuccessful or invalid experiments on instructional media. More specifically, it should, first of all, help those who wish to do experimental research to avoid at least some of the more obvious kinds of errors that others have made. It should thus help those who would conduct a successful experiment to have a reasonable probability of being, in fact, successful.

In addition, it should help even those who are never going to do an experiment to interpret and assess the so-called "experimental findings" in the literature of instructional-media experimentation. The reader of this literature will come across statements which say, in effect, that research "has shown conclusively" that factor X makes a greater contribution to educational films than factor Y, or that factor Z was shown to "make no difference." One of the things that we would like the reader to add to his intellectual arsenal is some basis for appraising such statements. He should know that, in some instances, from the very form of the statement it can't be true in the form stated; that in other instances it *might* be true but can't be established on the basis of the kind of evidence presented; and that in a few other instances the claim is indeed essentially supportable (even if perhaps stated somewhat over-strongly), at least within the limitations of the present state of the art. In any case, he should have some basis for knowing what kind of questions to ask of a research report (or reporter) in order to ascertain whether an alleged conclusion possibly is, or almost certainly is not, defensible. For it is unfortunately the case that such science of instruction as we have is at present in a state in which one must be extremely cautious about accepting at face value most "conclusions" derived from educational-media research studies.

It is important to be able to assess statements that purport, sometimes wrongly and sometimes rightly, sometimes knowledgeably but frequently ignorantly, to support rather sweeping conclusions with rather insufficient data, and to also distinguish those instances in which something has been reasonably well established. As we shall see, difference is not necessarily very much related to how many "cases" were tested, or to a number of other kinds of "requirements" which are frequently asked of experiments. Rather, it depends largely on some fundamental considerations in the logical design and analysis of the experiment, and in the definability or reproducibility of the "treatments" that were experimented with.

## Delimitation of Scope

This book does not deal with anywhere nearly all of the questions that one can raise about instructional media or all of the kinds of research related to media of instruction. Rather, as noted above, it is limited to research in which the effects of media are determined by experimental methods. As we shall see shortly, its scope is still further limited to dealing primarily with experiments designed to contribute to a science of instruction by testing hypotheses about causal factors that determine these effects.

The rest of this chapter and the next two chapters serve primarily to introduce and delimit the field of experimental research on educational media with which the book is concerned. We shall need first to say what we mean by the effects of media and to contrast questions concerning effects with other kinds of questions that one might ask.

### Exclusions

To begin with, let us identify certain kinds of questions that are important in the development and use of instructional media, but with which we shall not attempt to deal here. Among these kinds of research connected with educational media are studies dealing with problems of distribution logistics, curriculum analysis, ways of increasing the use of educational media, and library "research" of the kind customarily carried out as a basis for planning the content of instructional films or other instructional instruments. Such library research or source-material collation is an important step, of course, in the planning of educational instruments; we do not mean to minimize its importance, but merely to exclude it from the research on effects which is our primary concern here.

In addition to excluding such looking-up and verification of factual material, "survey" studies such as those to find out the current status of audio-visual aids or their use, or the consensus of teachers as to what kind of visual aids they would like to have available for teaching a particular subject, are also excluded. Similarly excluded are so-called "content-analysis" studies used to analyze the coverage of existing examples of communication media, sometimes conducted for the purpose of divining the unstated objectives of their authors. Finally, excluded also are many important questions that may contribute to experimental research on the effects of media, but do not deal with the nature and conduct of this research as such--for example, the problem of training well-qualified individuals to do research in this field, or selection techniques for picking promising research personnel.

## Effects of instructional media

We have said that we shall limit ourselves to studies which investigate the effects of educational media. Two further questions arise: What do we mean by effects? And how can the effects of instructional media be determined?

Effects as "learning" or gain. As a first approximation, the notion of the effects of educational media refers to the idea that students who have been instructed (through the use of media or by other means) are expected to be different after the instruction from what they were before. Thus, we are dealing with the effects of media on learning by students. "Learning" is, however, used here in a very broad sense, and could include any change in behavioral tendencies or capabilities as a result of experience. Thus it could include the modification of attitudes, prejudices, beliefs, convictions, motivations, and social habits, as well as the gaining of knowledge, skills, concepts, understandings, and the other kinds of outcomes which might be thought of as "learning" in a narrower sense.

Kinds of educational effects. Most studies of the effects of instructional media have been concerned primarily with effects as measured in terms of increase in knowledge or skill; however, a good many studies have also measured the effects of instructional media on opinions, attitudes or beliefs (e.g., Hovland, Lumsdaine, and Sheffield, 1949; Hovland, Janis, and Kelley, 1953). Some studies also have dealt with the effects of films on "interest" or motivations of students as evidenced in their tendency to undertake desired kinds of activities--for example, the study by May and Jenkinson (1958) on the effects of alternative versions of a film in stimulating students to read the book on which the film was based, or the study reported by Gropper and Lumsdaine (1961) of the effects of alternative television presentations in stimulating students to undertake experiments like those they had seen demonstrated in a televised science lesson.

These kinds of instructional effects are, in any case, to be distinguished from "effects" that are generally considered relevant in the "success" of a theatrical film or entertainment television program.

Here the success is assessed in terms of the popularity of the presentation as entertainment, in terms of box-office receipts or of indices of audience size such as the various TV and radio "ratings," or of the favorable or unfavorable notice given by reviewers. The success of instructional media, unlike that of entertainment media, cannot be assessed merely in terms of its popularity, or of judgments about its artistic merits, or of popularity, or of sales receipts. Rather, the criteria are primarily instructional effects assessable in terms of what students learn.

As indicated above, however, "learning" has to be broadly defined since, within the broad framework of education, there is a wide variety of purposes. The main purposes for a large proportion of instances in the case of instructional media may, however, be classified under the headings of imparting factual information, developing concepts or skills, and modifying attitudes or interests. It is important to note that, although a particular educational instrument may have more than one of these purposes, its success may be examined separately with respect to each kind of "effect."

Intended vs. unintended effects. In identifying the effects of instruction, we should also distinguish between intended effects versus effects which are adventitious or unexpected. Generally speaking, if instructional objectives have been defined carefully, the planned or intended effects would be generally synonymous with "desired" effects of instruction. However, this is not invariably true, because sometimes instruction produces unanticipated effects which are of importance. For example, a film designed primarily to teach certain physical principles might also have the desirable effect of increasing students interest in following a scientific career or the undesired side-effect of discouraging their interest in science.

Obviously the effects that are of interest in relation to a communication are generally related to its purpose. While we might properly look for indirect attitudinal or motivational effects, of a televised film series on physics or probability theory, we would certainly not fail to include among the effects to be assessed the more direct learning outcomes in terms of knowledge of the subject matter. On the other hand, the effects of a communication designed primarily for persuasive purposes would be assessed primarily in terms of the extent to which the communication actually swayed the audience in the direction of greater adoption of the belief, attitude or opinion being advocated, and only secondarily in terms of the related facts that might be learned from the communication in the course of presenting it.

An exclusion which involves a further delimitation of the field of this book is that we shall not deal very extensively with the results of research on purely persuasive communication--that is, those designed exclusively or primarily to change attitude or influence opinions. For the extensive studies in this area by Hovland and his co-workers, the reader is referred to the relevant portions of Hovland, Lumsdaine, and Sheffield (1949), and to later works by Hovland, Janis, and Kelley (1953), Hovland et al. (1957), Hovland and Weiss (1951), and others.

However, methodological aspects of such studies will be discussed as relevant to more general problems in methodology. Nevertheless, the exclusion from systematic coverage of the literature on communication and persuasion from the review of literature, presented in Chapters 13 and 14, implies a narrowing of the field of instructional media to concentrate on the effects of instructional instruments designed primarily to teach knowledge, skills, concepts and information rather than to mold opinion through persuasive communication. Obviously the separation of these two areas cannot be a complete one, and some studies are referred to in which interests and motivations or opinions were included among the kinds of effects measured.

*Measurement of effect*. Let us now look briefly at the ways in which educationally relevant effects can be determined. Here we are led inevitably to a problem of measurement. For example, if the effects of a film in imparting information are to be determined, we have to decide some way to measure these effects, and also have to decide quite explicitly, in terms of the film's purpose, what information the film will have to teach *in order to be regarded as successful*. We also have to decide on what basis we will attribute changes in information to the effects of the film. The term "measuring instrument" can be used here in a generic sense. It is usually a test of some form, but may include achievement tests, performance tests, or questionnaires or interviews for measuring effects on opinion or attitude. The question of measurement will be further discussed in Chapter 8.

Before-and-after measurement. Assuming for the moment that we have an appropriate measuring instrument, the essential requirement for measuring experimentally the effect of a film (or other instructional program) involves a fairly straightforward procedure. The measuring instrument can be used to determine, first, what the students who are to be instructed by it are like initially. For example, in the case of the factual film, this would be a test of how much they know about the subject matter before seeing the film. Second, the measuring instrument is used again to find out what the students are like--in this case, how much they know about the subject--after they have seen the film. The difference between these two measurements can then represent (subject to certain conditions which we shall discuss at some length presently) the effects of the film. The difference tells us, in other words, how much students have changed (learned) as a result of the film presentation.

A similar "before and after" application of other kinds of measuring instruments can be used regardless of the kinds of effects in which we are interested. Thus, for example, in the case of determining the effects of educational media on attitudes, an attitude scale of some kind which can measure the extent to which attitudes or opinions are held would be used instead of the test of factual knowledge; in the case of films designed primarily to arouse interest in a subject, techniques must be devised which can give a reliable assessment of "interest," or of the actions which indicate interest, both before and after the use of an educational program. In addition, as will be elaborated at considerable length subsequently, the determination of effects often involves the introduction of "control" measures designed to rule out erroneous explanations of the changes from "before" to "after"--i.e., changes that could reasonably be attributed to causes other than the effects of the film (or other educational program) under consideration. (Differences "with" and "without" having had an experience also represent one way of looking at the identification of an effect. But in this case also we have to exercise caution to insure that the "with" does not introduce other differences than those implied.)

Although the basic concept of "effects" as differences between what people were like "before and after" an experience is one which can be readily grasped in general terms, it should be noted that not all experiments which measure effects necessarily obtain before and after measures, at least not on the same individuals. Other ways in which this can be done will be noted in later chapters. We shall also discuss instances in which merely the difference in the "after" status of individuals who have been exposed systematically to different experiences (or "treatments") can tell us the difference in the effectiveness of the treatments, even though the absolute gain due to each was not necessarily measured.

Can all educational effects be measured?

In this connection a comment is deserved concerning the not infrequently encountered statements that some kinds of effects or outcomes inherently can not be measured. Our position is that, to the extent that this proposition must indeed be accepted, those kinds of effects might just about as well be forgotten about. If an effect can not be detected--if we cannot tell whether the kind of outcome that we seek has been achieved (either through our efforts or through some prior source of influence) there is little we can do about it.

This does not necessarily mean that we can measure all effects with equal dependability or precision or validity. Neither does it mean that it is necessarily easy or even generally feasible to measure all kinds of effects. Obviously, within a relatively short-term experiment we cannot measure effects which are only going to be revealed much later--say, after a period of twenty years. Certain of such effects might be called imponderable, and in some sense might include the possible effects of every action currently taken on the total density of future history of the world. But when people say effects cannot be measured, they are usually thinking in more limited terms than this, and often are speaking of such outcomes as effects on attitudes, motivations, interest, appreciations, and the other "broader" objectives of education. (Sometimes, too, there seems to be the implication that only these effects are really important, or the implication that effects other than these are necessarily limited to relatively unimportant simple rote mastery of facts or other rote associations.)

Although this issue does not arise explicitly in respect to many of the considerations with which we must deal, we should record our disagreement with these general positions. In the first place, we believe that the imparting of specific knowledge is, in itself, while by no means the only or necessarily even the most important kind of educational objective, still one of considerable value. Facts are helpful things--let us not despise them; simple skills and the ability to manipulate certain relationships are valuable too, even though elementary. In the second place, we would also disagree with the notion that there is something inherently unmeasurable about attitudes, motivations, interests, appreciations, etc. (Indeed, considerable success has been attained in defining these in such a way that they can be measured and in going about measuring them. To observe this, one has only to consult some of the better studies of attitude change.[*] ) The difficulty is, usually, merely in getting agreement as to an operational definition of the "broader objective"--e.g., of translating "interest" into specific kinds of behavior which we will agree are indicative of what we mean by "interest," or of agreeing on the observable behavior which will differentiate between a "motivated" and an "unmotivated" individual.

A final consideration with respect to "measurability" relates to the essential notion of measurement. The position that we shall take here follows the dictum that everything which exists, exists in some degree, and therefore can, in principle, be measured.[**] It seems to us that this position follows for any characteristic, achieved by instruction or otherwise, which can be detected, or observed as presence or absence (or presence to a greater degree or to a lesser degree); and we have already indicated that unless this is true, the potential outcome has little practical significance--or, at least, is not amenable to inquiry at other than the level of the wooliest of armchair philosophizing. The measurability of any quality which can be detected follows directly from the realization that in its simplest form measurement reduces to counting. Hence if the presence or absence (or any degree of difference in the presence or absence) or any educational outcome of potential interest can be detected or discriminated, we can at least count the number of individuals or instances where it is present or absent, stronger or weaker, etc. In this limited sense, at least, we can surely then measure appreciation, motivation, attitudes, beliefs, etc., regardless of whether we are able to measure them with a degree of precision (or even define them with a degree of refinement) which we should like. A relevant comment here is that of Wallis and Roberts (1959, p. 7), who say:

> Often, of course, summarization of important observations must necessarily be "impressionistic" or "literary" rather than numerical; this is true, for example, of anthropological studies of the character and values of cultures, or of art criticism. The statistical approach is limited to those aspects of things that can be described and summarized numerically. This limitation is not, however, as confining as it may at first appear. Many things that are "qualitative" or "subjective" nevertheless have a quantitative aspect; for example, an important aspect of a certain organic disease may be the number of times it occurs. Many subjective or qualitative impressions can be sharpened or corrected by statistical study of subsidiary details, as when the impression that racial discrimination is decreasing is checked against the number of occurrences of certain specific kinds of incidents.

Non-experimental approaches to identifying effects

Regardless of the kind of effects measured, the actual measurement of differences that are produced as a result of exposure to or use of educational media implies what we will term an experimental approach to the determination of media effects. Although attempts are sometimes made to assess effects by other than experimental means, our attention will center almost exclusively on the use of experiments for this purpose. Considerable elaboration on what is meant by an experiment will follow in succeeding sections. For the moment it will suffice if the term is understood to mean merely the determination of effects by determining changes due to controlled exposure to media, through some procedure that involves essentially the kind of before-after rationale earlier characterized.

Concentration on experimental methods means that we shall give relatively little attention to methods such as the collection of reviews and "evaluations" from preview groups, or the results of surveys that infer effects from essentially correlational data--that is, from testing the knowledge of beliefs of individuals who merely *happen to have been* exposed (or who report having been exposed) to a particular educational-media presentation, as compared with those who happen not to have been so exposed. (See further discussion in Chapter 3). For our purpose, the exposure or non-exposure must have been *determined arbitrarily* in accordance with an experimental procedure.

Similarly, we will not be dealing with attempts to infer or predict effects of media on the basis of expert judgments or educated guesses, nor from interview data in which individuals report how they think they were themselves affected by educational instruments. The reasons for placing almost exclusive dependence on the results of *experiments* as an index of the effects of media will become clearer when we consider the nature of experiments and the kinds of fallacious inferences which a carefully controlled experimental design is intended to preclude.

A simple experiment. The simplest form of media experiment might ask merely whether a particular film (or other instructional program) does or does not produce certain effects. Generally this is not an especially interesting question (though it may be an important one to ask before proceeding with more elaborate experimentation, in order to assure adequate conditions for testing more interesting hypotheses effectively: (see Chapters 12 and 15). We will return presently to somewhat more interesting questions involving comparison of alternative methods, or forms of an instrument, to test hypotheses about factors alleged to make one form more effective than another.

For the sake of specificity, consider the effects of a film as an example. For "film" you can read "T.V. lesson", "self-instructional program" or "language laboratory lesson" if this makes the discussion more relevant to your interests. We will also assume, for specificity, that our interest is in whether the film conveyed actual information about its subject matter--which might, say, be the geography of Tibet. You can substitute a different objective if you like, without making any difference in the argument; what we have to say would apply equally if our film (or other instrument) were instead intended to teach a skill (e.g., use of a slide rule, or trouble-shooting a T.V. set, or solving quadratic equations, or tying a bow-line), or if it were to develop interests or attitudes (e.g., getting students to want to read a book, or to do a physics experiment; or getting them to accept the idea that school segregation is undesirable, or that experiments are difficult to do properly).

Suppose we merely want to test whether the film on Tibet had any effect. The "independent variable" which we manipulate is presence or absence of the film. To do this in our experiment, we arbitrarily withhold the film from one group and administer it to another equivalent group. Then we test both groups. For the moment, we won't raise the question of whether or not this is a good experiment or a worthwhile experiment, but merely point out that it is, basically, an experiment. On the other hand, suppose we ask each of a number of people whether they have seen the film, and then compare knowledge about the subject for those who say "yes" versus for those who say "no". This we would not call an experiment. The reason that it is not an experiment is that we did not administer the independent variable arbitrarily; rather, we took it as we found it. We took those people who said they had seen the film, versus those who said they had not seen it.

Even if we somehow know their reports are correct (which they of course may not be) their having seen it or not was due to some cause that was completely beyond our control. If we know that difference in knowledge between the "saw" and "didn't see" are statistically reliable, we can refute the null-hypothesis that both groups are equally knowledgeable. But this does not necessarily mean that the difference in knowledge was taught by the film. Why did the "see" group see that film? Perhaps because they were interested in the topic as a result of having studied it previously. Maybe they didn't learn a thing from the film, but simply were knowledgeable before they went to see the film. All we know is that one group includes those who had seen it and that it also includes those who knew more. But we do not know that the seeing of the film was the cause of that greater knowledge. Related considerations will be stressed at several points in our further discussion of experimental logic and design.

"Control" in an experiment. An experiment to determine whether seeing our film resulted in students' learning any information about Tibet will require, like any other experiment, certain elements of control. What this basically reduces to is a way to rule out alternative explanations of effects that might otherwise be attributed to the independent variable (here, presentation of the film). The concept of control and its application in experiments in one form or another will enter into each of our subsequent discussions.

In excluding non-experimental bases for the determination of effects, we are not defending our choice by taking the extreme position that cause-effect relationships can never be established except through experiments. Consideration of the status of astronomy as a science--despite the inconvenience, to say the least, of trying to introduce arbitrary manipulation into cosmic events--is sufficient to preclude our wanting to adopt this extreme view. For example, Tukey (1961, p.119) points out, in connection with the crucial role of randomization in experiments, that:

> In truth, randomization is not only one of the major keys to the establishment of causality, but a great aid in the establishment of empirical fact. Yet astronomy, to name an ancient and honorable science, has learned much with little randomization.

Limiting the book to "experimental research" does not even mean the authors think it is the only type of research of value or the only kind that needs doing to improve education. It does mean that we feel experimentation is the method of choice for building a science of instruction, for getting answers to questions about the effects of media and (less precisely, perhaps) identifying the causes of these effects. In the history of media research on effects, moreover, non-experimental approaches have generally been of relatively little worth, and the problems attacked could better have been treated experimentally. Further, we feel that experiment provides a better model for the understanding of research than do other methods, and serves as the best point of departure for considering other possible methodologies (see further discussion in Chapter 3).

It may be noted that the question of the relative utility of experimental versus non-experimental research is often raised in connection with the social sciences, and that even as rigorous an intellect as Tukey does not place so exclusive an emphasis on experiment, as a tool for the social sciences generally, as we do here in the particular case of instructional media. For example, Tukey (1961, pp. 119-120) says:

> There may be some who question "experimentation" in social science on the ground that social sciences are observational rather than experimental. Comparison of the possibilities of directed, comparative observation with those of experimentation, shows that, while causality is far more easily established by experimentation, subtle empirical facts can often be almost as easily established by comparative observation of situations *selected* to involve properly related environmental variables, as by comparative observations of situations *arranged* to involve them. . . . Social and behavioral science needs to learn how to do more experiments with a capital E, experiments which include randomization and all the other precautions, but it cannot afford to confine its uses of the principles

of wise experimentation to such exceptional instances.

The further reason for our exclusive concern with experiment in this volume, despite the cogency of the last part of Tukey's advice to the social scientist, lies in the special properties of reproducibility and realistic "introducibility" that are characteristic of instructional media as educational tools. In view of these properties, as was pointed out by the senior author some years ago (Lumsdaine, 1951), the conduct of well controlled experiments is much more feasible than is often assumed; thus, the frequency with which less satisfactory methods are resorted to is symptomatic of an unnecessarily defeatist attitude on the part of investigators. We would continue to argue that, at least in the case of instructional media, more effort should be devoted to capitalizing on available experimental possibilities, and less effort to rationalizing about the difficulties of doing experiments and to attempting to squeeze causal inferences out of non-experimental data.

Obviously, however, the word magic of describing a study as an "experiment" does not automatically guarantee sound results. Sloppy execution has perhaps made as many experiments worthless as has improper design. Experimental techniques in the field of films and other mass media are being improved, but wider knowledge of the use of good experimental methodology is needed. An example in experiments is the fact (also pointed out some years ago: Lumsdaine, 1951, 1953) that in using group sampling and group administration, as in the case of films, it is generally the number of independently sampled groups rather than the total number of individuals tested that should determine the degrees of freedom for estimating error of measurement. Nevertheless, this requirement--further described in Chapter 10 --has been frequently ignored in educational-media experiments that have employed group sampling.

Purposes of Experimental Research

Assuming that we can measure the effects of an instructional instrument by some form of experimental procedure such as controlled before-and-after measurement, we can raise the question of the purpose to which such determination of effects can be put. The first of these purposes can be called basic, scientific, or "hypothesis testing" research. Its essential purpose is to add to our knowledge of cause-effect relationships as a basis for better design and use of instructional media to the development of a science of instruction. The second purpose applies to what can be called "applied research", "production research", "technological research", or "evaluative research"; however named, it has as its proper purpose an immediate and highly specific practical goal: for example, deciding which of two existing films to use, or obtaining data to know what effects were produced by a particular film or self-instructional program as a basis for revising and improving it, or as a basis for knowing what supplementary instruction might be required.

The distinction between research appropriate to these two purposes will enter into the discussion in several places in the book, particularly in Chapter 12 in discussing classes of experiments in relation to the choice of research problems. However, a brief elaboration of the distinction here may help as a basis for understanding the essential difference between the two. For these purposes we may take motion picture films as an example.

Using films as an example, technological (or production research or evaluative research) would ordinarily represent the measurement of the effects of a single film. Such a study can tell the producer of the film which points of his presentation are getting across and which ones are not. This information is of particular value if carried on in a preliminary version of the film prior to its completion and release, so that the results may be used in correcting or reinforcing weak spots in the film's content. As just suggested, such data could also be used if obtained after the completion of the film as a basis for designing supplementary materials to be used in conjunction with it. Research of this kind has the advantage that it can be carried out relatively rapidly and the results can, in principle at least, be utilized immediately. It has the major limitation, however, that data obtained about the effects of a single film are largely restricted to their use in improving or assessing the particular film studied and contribute little in the way of generalizable knowledge to help in the design of future films. The same limitation evidently applies in the case of other instructional media--e.g., self-instructional materials--that is, the results of an evaluative study to measure the effects of a single program are limited to assessing or improving that particular program. We shall have occasion to comment on the reasons for this limitation more than once in subsequent chapters.

The other broad classification of research, by contrast, is directed toward obtaining generalizable knowledge by comparing effects of two or more instructional instruments which differ with respect to specified characteristics--for example, a film which presents a subject matter in two or more different ways. Potentially generalizable conclusions can thus be generated about methods or techniques of presentation, particularly (sticking to films as an example) when experimental comparison of effects is made for two films which *differ only with respect to a single specified factor but are* alike in all other respects. Such comparisons may be used to test hypotheses about effective methods of presentation--or, more generally, effective methods of instruction--and thus contribute to principles which can be used in making decisions about the design or use of future instructional instruments. It is to such principles that we refer in speaking of a potential *science of instruction*.

What has just been said is necessarily an over-simplified and very incomplete account, but it may serve temporarily to identify the main differences in purpose and requirements for the two broad uses to which the measurement of the effects of an instructional instrument may be applied.

In the "technological" use of experimental techniques to measure the effects of a single, specific educational instrument, we might say that a hypothesis being tested is merely that the instrument has, in fact, certain effects ( as against a competing null-hypothesis which would assert the instrument has *no* effect on the measure or measures in question). However, in such cases we would, characteristically, be interested not just in a single effect, but in the effect of the instrument on a variety of specific sub-goals--e.g., specific points of information, specific convictions, beliefs, opinions, specific components of a skill. Although we might wish to assess the overall outcome of a single combined measure, in doing so we would normally be interested in more than merely rejecting the hypothesis of no effects at all of the instrument, since presumably any fool can make an educational presentation which will at least have _some_ desired effect on the audience (though even this is not certain; or, at least, it is not certain that sometimes undesired effects may occur which more than outweight desired effects).

In the case of such an "evaluative" or "technological" experiment, furthermore, we are likely (to a greater extent than in the case of scientific experiments) to really be dealing with a problem of estimation--that is, it becomes quite important to know not merely that an effect on a particular desired outcome has taken place, but to get some fix on the _size_ of this effect.

Similarities in the requirements for experimentation of the two types

The similarities and differences between the two above identified classes of experiments, or classes of experimental research, will be returned to several times in the present volume. As we shall see, the distinction is a very fundamental one, which enters into such diverse problems as the choice of what research to do, the kind of experimental design to be employed, the degree of methodological rigor that we feel should be insisted on, the setting of "confidence levels" or "rejection levels," etc. The distinction will appear in a number of guises--"technological" versus "scientific" experiments, "applied" versus "basic" research, decisions versus conclusions, product testing versus hypothesis testing, "evaluative" experiments versus "analytic" experiments, generalizable versus non-generalizable hypotheses, and others.

On the other hand, as is also pointed out elsewhere (see Ch. 3) although the requirements for the two classes of experiments differ in certain very important respects, the methodological considerations involved are often very similar and, in some respects, identical. The latter is true, for example, with respect to certain problems of measurement, internal validity, sensitiveness--or at least precision, and efficiency of experiments. Furthermore, some of the basic characteristics of experiments and the problems that enter into them can most easily be discussed in the general case of merely comparing Treatment A versus Treatment B, without reference to the scientific versus technological purpose of the experiment--without reference, that is, to what it is that is being compared. In some cases, examples of certain distinctions important in the structure of an experiment can be found more readily in the case of an experiment having a specific "technological" purpose (like comparing two specific films or programs to see which one we wish to adopt). In most instances, however, we have chosen our examples from scientifically oriented experiments, directed toward the testing of a hypothesis of some potential generality.

Method of comparison considered without regard to "what is compared".

A separation of what is compared (or what is hypothesized) from how it needs to be compared (i.e., what are the formal requirements for a valid comparison of the two treatments) cannot be a complete one;

to some extent the two kinds of questions interact. But in general, most of our discussion for the next few chapters will apply regardless of the specific content and degree of generality of the hypothesis, provided only that it unambiguously specifies what things are compared. Thus, we might compare Mr. Brown's film with Mr. Smith's film, or two unidentified programs marked A and B, with no specification of how the content or organization or characteristics of the two differed, and still profitably discuss at considerable length--as we shall do--what kinds of things would be necessary to make a valid comparison between the effects produced by A and those produced by B, without further looking at the question of in what respects A and B differ.

Restriction to hypothesis-testing experiments

Of the purposes to which experimental measurement of the effects of instructional media may be directed, this book will be limited primarily--as has been already implied--to experiments directed to establishing principles or testing of hypotheses. Essentially this means that a prediction as to which of two alternative forms of an instructional instrument will be superior is tested by comparing two alternative versions of an instructional instrument. One version exemplifies the method hypothesized to be superior, and the other version exemplifies an alternative kind of instruction implied by the hypothesis.

The forms of experimentation with which we shall be concerned are primarily suitable to determining merely whether a hypothesis about the advantage of one form of instruction is tenable--that is, whether the evidence does or does not support the hypothesis. This is not the only kind of experimental question that can be asked. Obviously, it would be desirable to know not only *whether* one method is superior but *how much* superior it is, or to be able to assign a measure to the degree of difference in outcome that results from a specified kind of variation in the instruction. Such questions fall within the sphere of what can be called "*estimation*" or ("estimation of parameters") and, perhaps suprisingly, will not be dealt with very extensively here.

The reason for this further delimitation lies primarily in the difficulties of making dependable and valid estimates of this sort. This in turn is due in part to the limitations in our present ability to measure human performances or the outcomes of instruction in units that have generality from one situation to another. In any case, most experiments that have been done on instructional media are, in fact, at most able to demonstrate only whether one alternative is superior to another without definitely saying (in any generalizable terms, at least) *how much* superior. In other words, most media experiments *are* hypothesis-testing experiments (even though frequently not recognized as such), rather than experiments that provide a basis for estimation of the magnitude of effects, or of the absolute difference in magnitude on the effects, produced by two different instructional programs.

*Depth of coverage: Implications of the word "guidebook"*. As noted previously, this book is not intended as a comprehensive textbook, certainly not as a self-instructional program. It is a sort of reference book, but still it is not a handbook. It is not a "how to do it" book, like a cookbook. It presents information which the prospective neophyte in research should first study, and then should make frequent reference to, at major points of decision in planning and carrying out research. It is also not a self-contained book; for example, it does not cover comprehensively the details of statistical technique that are closely tied to certain kinds of formal aspects of design. Rather, in the main, it relegates statistical *details* to appropriate reference sources where detailed information on the relation of formal design and statistical treatment can be located. Similarly, though some specific experiments are described and the results of others are summarized, the authors have not felt it feasible to encompass within the present covers the range of illustrative experiments to which the reader should have access. Reference to a few primary sources of such illustrative experiments is therefore assumed, as is reference to one or more basic statistical texts.

The content of this book, as has been indicated, deals with matters that the authors feel a person who is doing or planning research should be thinking about and have in mind. It is to be stressed that the authors believe *some of the more elementary-sounding considerations are of basic importance*--that is, they are crucial considerations, *not* at the "i-dotting and t-crossing" level. This is true even in a few instances where, because of experience in the application of research design to the special problems that are encountered when research is done with educational media, the position the authors take differs somewhat from the conventional treatment of experimental and research design. Important issues that seem crucial in experimentation will be raised even where, in some instances, the authors do not feel we have satisfactory answers to them. In this case, we will try (in a supplementary appendix) to identify more precisely the nature of the impasse or problem to be solved which appears to need methodological advances before a good answer can be given.

# Chapter 2
# Educational Media as Objects of Research

## Education, instruction and media

Since this book deals with "educational media", we should perhaps be more specific about which media we are going to discuss. The term "instructional media" may be substituted for "educational media", since our primary concern is with the effects of using educational media in attaining definable instructional goals. 'Educational media" refers to media for presenting or administering an education "program" or lesson in a sequenced, pre-programmed form. Specifically included are films, televised instruction and "programmed instruction" (the last means programmed in the broader sense as well as in the narrower meaning ascribed to programmed "self-teaching" or "auto-instructional" materials used in a "teaching machine" or programmed text"). In general, "educational media" is used herein more or less interchangeably with "instructional media", though sometimes perhaps with a slightly more specific connotation indicated in the word "instruction" as compared with "education". By instruction we mean any form of teaching intended to achieve a definite specifiable educational outcome. Instruction thus has a connotation somewhat less broad than does "education" because of the emphasis on quantitatively specifiable objectives or outcomes being sought (which might not be true of some of the more diffuse aims of education).

The word "instruction" is here used as a generic term, referring to any specifiable means of controlling or manipulating a sequence of events to produce definable modifications of behavior through learning. We are dealing only with those outcomes of learning which can be specified in sufficiently explicit terms to permit their measurement. As noted in Chapter 1, these outcomes may include changes in attitudes, interests, motivations, beliefs or opinions so long as these are defined in terms which identify specific, observable behaviors agreed to be manifestations or indications of these outcomes. Use of the word "outcomes" rather than "objectives" does not, however, preclude considerations of *any* outcome--whether anticipated, desired or otherwise--even though our main concern is with the attainment of objectives which are chosen in advance to represent the purpose of the instruction (cf. Chapter 1, pp. 17-18).

As used in this book, "*instructional media*" refers either to a particular class of instructional resources (e.g., films), or to a group or subset of such a class, or to media collectively. An instructional "*instrument*" or "*program*" generally refers to a particular member or instance of this class, e.g., a particular film, self-instructional program, televised course, etc. The terms "instrument" and "media" can refer either to the instructional materials prepared for teaching a particular subject matter, or to the combination of equipment and associated materials--e.g., a film and projector, or a teaching machine and the program materials presented by it.

Instructional resources which can aid in the learning process include, in addition to the teacher, a wide variety of tools, devices, materials and other "aids". Broadly speaking, these either present, or help to present, information to the learner, or afford opportunities for him to make responses which help him to learn or indicate what he has learned. The "audio-visual" media of instruction of course comprise an important and widely used class of such instructional resources, generally employed to present information. Other instructional tools include instruments which permit the student to make relevant responses (e.g., workbooks, self-scoring test forms and other manipulanda) which can aid the learning process. The use of "programed instruction" materials, particularly teaching machines or "auto-instructional" devices, serves both an information-presenting function and also affords a means for the student to make appropriate responses to the material presented.

The term *programed instruction* means, generically, any presequenced instruction given by instructional media to groups of individuals. More specifically, depending on context, this term may refer more narrowly to forms of programed instruction (or "programed learning") which use teaching machines or programed-text material to present individual instruction to a single student who is learning at his own rate, with the additional characteristics of requiring frequent response by the student and giving him immediate feedback (confirmation or correction) for each response he makes. Note that one of these characteristics can be present without others; in particular, frequent student response and immediate feedback may be employed in group instructional media by films or television.

Programed individual instruction further implies an individually determined rate of presentation and response and permits the student to proceed for an appreciable time without the intervention of a teacher. This latter characteristic is also true of many forms of audio-visual media, e.g., sound films, in which the student is supposed to learn from the audio-visual presentation without the continuous intervention of the teacher, just as he is expected to do when pursuing independent study of a textbook or reference source. Some other visual and audio-visual "aids" by contrast, are used by the teacher as adjuncts to his other activities, rather than as independent devices of instruction. (All instructional instruments and the media they present can properly be considered "aids" to learning. However, not all need be considered aids to the teacher, since some, e.g., self-instructional programs, or in a sense, textbooks, are clearly capable of instructing without the teacher's direct intervention.)

## Educational Media as sequenced and reproducible presentation

The educational media with which this book deals are primarily those presented in some predetermined sequence. The sequence is determined either by the structure of the instructional instrument, by well defined procedures and constraints in the way they are used, or by the mechanical arrangements used to present them. The instruments thus generate events whose occurrence is more or less predictable in terms of sequence or pacing. The sequencing and pacing need not be fixed, but they do depend on clearly specifiable factors in the learning situation. A specifiable factor does not mean merely the activities of an instructor, nor the decisions of a student in whose hands materials are placed. This restriction does not mean that media presentations may not require the activity of a teacher or assistant (e.g., film projectionist); it does mean that the student is expected to learn something from a previously prepared sequence of instruction even though the teacher is not present or active at the moment. Instructional media, as the term is used here, provide acts of instruction, not merely the wherewithal for such acts.

The events generated by instructional media also tend to be relatively *reproducible* events in that they may be repeated (as in the case of a film) or be duplicated (as in the case of instructional television) at a number of locations simultaneously. In concrete terms, the instructional instruments dealt with here include motion picture films, tape recordings, filmstrips, and self-instructional programs for presentation by a teaching machine or some related device. Instructional programs or instruments do not, in general, include charts, diagrams, models or isolated photographs or drawings, as such, except when these represent elements in a predictable pattern of instruction governed by a relatively well defined procedural sequence. Such isolated elements can be regarded only as tools or components for instructional acts or instruments; a complete instrument consists of the aid plus the activities of the teacher. This restriction is necessary if we are to consider instructional media and instruments as specifiable events which are amenable to scientific research, which can deal only with reproducible events.

As the senior author has pointed out elsewhere (Lumsdaine, 1960 ) single visual displays represent only an isolated and incomplete fragment of a total instructional stimulus pattern. Thus, except when they are considered only as a component of other specified instruction, it is difficult to control or predict their practical effects on instruction. When they are considered alone, it is difficult to partial out their contribution from other unspecified factors in the instructional situation. Perhaps this is why little significant experimental research has been done on non-sequenced, isolated visual aids.

In terms of the criteria set forth above textbooks represent a borderline case. The usual textbook does not control the behavior of the learner in a way which makes it highly predictable as a vehicle of instruction or amenable to experimental research. It does not in itself generate a describable and predictable process of learner behavior, and this may be the reason why there has been very little productive experimental research on the textbook. In any case, the textbook and its cousin, the training manual or training syllabus, are not treated explicitly here as a primary object of research.

## Major Classes of Instructional Media

### Instructional films and related media

The terms "teaching films," "educational films," "training films" (the military and industrial term) all indicate presentation of material by motion-picture film as used for instructional purposes. Most instructional films are currently produced as 16 mm. sound films, in black and white or in color. They present an enormous variety of instructional materials, including film-recorded lectures, surrogate field trip visits (the travelogue and other films which aim primarily to "bring the world to the classroom"), dramatic or theatrical presentations of literary or historical events, and dramatized forms of exposition which depict the teaching of students within the film by a teacher or parent. In addition, two quite different kinds of "demonstrational" films should be distinguished: first, demonstrations of *phenomena* depicting events or characteristics of the animate or inanimate world and, second, *procedural* demonstrations designed to teach the performance of a specific sequence of acts.

*Instructional objectives of films and related media.* Aside from some instances in the latter category, sometimes deprecatingly referred to as "nuts and bolts" training films, the instructional objectives for which films were presumably designed unfortunately tend to be stated imprecisely, if at all; when not characterized merely as "telling a story about" a subject matter, or "presenting" the "content" of a lesson (without specification as to what precisely is to be learned), the objectives are often described as those of creating a general familiarity, understanding, background, or "knowledge about" the subject treated. Because of this vagueness of objective, it is often difficult to define the appropriate criteria for evaluating instructional films. That the same often applies to other, non-media forms of instruction is perhaps less often obtrusive because of the fact that non-media instruction is less often made the subject for experimental evaluation.

*Film production.* Most instructional sound films are now produced either with synchronized sound recorded concurrently or with narration and other sound added separately. Their high cost when commercially produced can sometimes be very greatly reduced by local, short-cut production techniques such as those described by Greenhill (1955).

Often the content and sequencing of instructional films are governed in part by practices and traditions borrowed from the field of theatrical motion pictures as much as by considerations of pedagogy.

__Films and kinescopes or video tape recordings__. The respects in which films are likely to differ instructionally from a kinescope or video tape recording of live televised instruction probably lie more in philosophy and practice of production than they do in the inherent characteristic differences of the media. (Television not prefilmed or recorded by video tape or kinescope is, of course, transitory and not temporally repeatable.) Aside from minor differences in grain or resolution (these differences are very considerable in terms of physical measurement but probably are seldom important from an instructional standpoint), the televised instructional lesson, kinescope recording, and video tape recording differ instructionally from the sound motion picture primarily with respect to screen size and in the potentiality for using color, except in cases where immediacy of presentation is a genuinely important factor. Even this difference is more a matter of tradition and usage than of inherent properties, however, since motion picture film can, by special equipment, be processed and projected within a matter of seconds from the time an event is photographed. For most instructional purposes, we shall therefore consider the inherent instructional properties of film and of recorded television presentations as substantially identical. Both consist, in fact, of a series of moving or sequenced still pictures accompanied by recorded sound, and in many instances only an expert can detect the difference between pre-filmed or taped and live television presentations, or between films photographed directly on a photographic emulsion and those in which the original action was picked up by a television camera.

Differences in production practices. At the present time, however, films and televised instruction often differ considerably because of tradition-engendered factors related to inherent or adventitious constraints in the budgets and time schedules for their production, and also because of conventional practices such as that of filming television-linked films "straight through" (in contrast to standard practices in producing films by shooting one scene at a time, and later putting the resulting bits and pieces together by a careful job of editing in the cutting room.) This difference in tradition appears to be weakening, however, with the increasing use of pre-filming for both instructional as well as entertainment television, and of the so-called "straight through" or "television-style" techniques. (The feasibility of the latter is somewhat enhanced by the so-called "telecam" technique in which the same image is picked up concurrently on film and over a video circuit, enabling on-the-spot direction and editorial decisions to be made, while at the same time preserving a high quality film original for subsequent reproduction.)

Leaving aside these questions of production technique, most of the variation in cost as well as instructional properties results from what is placed in front of the camera. It may also be noted (e.g., in the subsequently described Lumsdaine-Sulzer-Kopstein study: 1951, 1961) that the important characteristic of reproducible, carefully planned instruction, as represented by film and its TV cousins, can be instructionally valuable even where the visual material is largely of a static sort.

Similar instruction in this case can often be provided by a sound-accompanied series of still pictures, requiring simpler equipment and less expensive materials in some instances.

Semi-motion presentations. Some forms of devices for sequenced still-picture projection are not only capable of regulating the timing of successive frames or scenes by automatic signals from the accompanying sound recording, but also are capable of changing frames rapidly enough that simulated motion, pop-on effects, and other characteristics of motion-picture presentation are closely simulated. When one considers also the actual absence of essential motion, and frequent lack of need for it, in film sequences running many feet in length of motion-picture film, it is evident that the distinction between "film" and "filmstrip" media is a hazy and shifting one, clear only in the extremes and even then better specified in terms of specific stimulus properties of motion and transition than by characterization in terms of "media" properties.

For group-presented pre-paced (as well as pre-sequenced) presentation, it is true, however, that presentation by motion-picture film projector has maximum flexibility in the sense that it permits any combination of still and motion presentation. On the other hand, the motion picture is sometimes at an important cost disadvantage where motion is not really utilized effectively for considerable periods of the instruction. In addition, flexibility is lost since, with most existing forms of motion picture projection equipment, at least the pacing or timing of the instructional material cannot be guided by the reactions of an individual learner. This has thus far not been considered a primary disadvantage because of the fact that film presentation has generally been used as a group medium. However, with the advent of teaching machines and the concepts associated with them, increasing use will be made of individually presented films having flexible start-stop characteristics.

Film utilization in education. It is difficult to obtain firm data on the prevalence of film usage; no exact counts are readily available on the total number of educational films in current use in our schools and colleges, nor on the total number of showings or pupil hours of film viewing presented per year. The total number of films available, however, is very large; the most comprehensive single catalogue, Wilson's lists over 10,000 films in the 1954-58 edition and subsequent supplements. Teachers have increasingly been trained and encouraged to use films, as well as the simpler visual aids to instruction; some school systems require teachers to have had at least a one unit laboratory course in the operation of film projectors and related equipment.

Characteristic guidelines for the use of films have stressed the role of films as aids to teaching, rather than as self-contained instruments of instruction. Teachers are encouraged to preview films, to provide introduction to the films for students, to conduct followup discussions, and less commonly, to intersperse question and answer sessions or other class participation exercises between sections of a film. Although these practices are unquestionably useful in many instances, the perspective of regarding the film merely as an adjunct to other instruction is by no means a necessary one, and there has been an increasing tendency to construct films which are intended to stand on their own feet, if necessary, as independent instructional media. A well-known example is the Harvey White physics series, which was initially filmed on the basis of a series of televised lessons. One suspects that the emphasis on using films as aids is one which is slowly changing as better and more comprehensive films become available, and that it has been, in the past, rooted partly in the knowledge that many films are ill-equipped to stand on their own feet as independent acts of instruction, and partly in apprehension lest the film displace the teacher. It is interesting to note that, in some instances, recent uses of films are perceived as carrying the main burden of instruction with the classroom teacher aiding the film presentation, rather than the film being used as an aid to a sequence of instruction organized by the teacher. An advantage provided by the "self-sufficient" televised or filmed lesson (shared with self-instructional programmed materials) is that it makes it possible to offer courses of instruction on a wide variety of subjects in small schools where qualified teachers are not available for all of these subjects.

Further aspects of television

Instructional television has tended even more than films to assume the primary role in conveying instruction; even where a teacher is present, her role is often a supporting rather than a supported one. Where complete half-hour or hour-long lesson periods are transmitted by television, the classroom teacher in some instances provides follow-up, or preparatory activities; in other instances, the televised instruction in a particular subject is the only instruction that the students get. Whatever disadvantages may be raised for this latter pattern, it represents, evidently, an increasing trend. However, closed circuit television has also been utilized as an aid to the instructor--for example, in the use of the overhead television camera as a tool at the disposal of the lecturer to replace slide projections and other forms of projected aids, or to transmit demonstrations, e.g., in science and medicine, including surgical operations, so that more students can get a good view of the demonstration. However, the primary role has been that of televising a complete lesson, and the increasing tendency to record these lessons--and the obvious wastefulness of not doing so when extensive preparation and resources have gone into their production--makes it further apparent that, for many purposes, televised instruction and equivalent instructional films are identical media. Certainly, research on factors that make for effective presentation of information on a film screen should, in most particulars, apply to equivalent instruction projected on a television screen.

Broadcast and closed-circuit TV instruction. In addition to the "educational" programs transmitted over television for public information or general enlightenment, many class hours of instruction are being transmitted over educational stations in the United States. Approximately 50 stations are licensed at the present time as educational stations on UHF or VHF channels, and many of these devote considerable time, particularly during the school day, to televised lessons for use in school classrooms or as extension credit courses. For example, WQED in Pittsburgh and its UHF affiliate, WQEX, together currently transmit over eight hours a day, including lessons in subjects ranging from shorthand to physics and providing basic course content for approximately ten courses in addition to supplemental programs for numerous other subjects. Extensive coverage is also provided by an increasing number of closed-circuit installations operated by school systems (e.g., Hagerstown, Maryland, and Anaheim, California). The advent of an airborne television facility designed to blanket a wide geographic region with televised courses further dramatizes the increasing utilization of transmitted instruction and emphasizes the importance of research on presentation characteristics.

_TV as an aid_. As noted above, the use of television is not limited to transmitting an entire lesson or series of lessons over a closed-circuit or open-circuit (broadcast) channel; it may also be employed as an aid to an instructor in a classroom. Bretz, at the University of California, Los Angeles, has developed a portable installation which can be quickly set up in most classrooms and which employs an overhead camera at the instructor's podium to substitute for a variety of other projection facilities and to eliminate the need for special transparencies. Textbook illustrations, charts, drawings, and other visual material which the instructor wishes to use in a lecture are simply placed on the desk under the camera and are projected on one or more monitor screens in the classroom. The use of a zoom lens on the overhead camera permits adaptation to materials of different sizes by increased magnification for showing detailed portions of a display. (See Wade and Bretz, 1960.)

### Audio instruction

Unquestionably the major current use of audio instruction in most schools in the United States is for teaching foreign languages. Although a number of recorded lectures have been made available to schools, their use and that of radio-presented instruction has never caught on in the United States as the use of films and TV-presented instruction has. Thus far, less attention appears to have been paid to making well-designed instructional recordings available than to the installation of language laboratory equipment, the assumption evidently being that the teacher would prepare materials locally for use with the equipment.

Hundreds of such language laboratory installations have been made, generally including a magnetic recording system employing tapes or, occasionally, discs. Students listen to recorded material in the foreign language and then attempt to reproduce it. A number of the systems allow recording of the student's effort so that he may compare it with the master recording he has imitated. The effectiveness of such instruction depends not only on the availability of recorded materials, but also on the characteristics of the equipment. In some instances, this permits the student to proceed at his own pace, using an individual recording. In other instances, the group is lock-stepped to a common master recording. In some systems, the student is able to play back a comparison of his own vocalization efforts *immediately* in comparison with the model; in other systems, playback and comparison has to be delayed for an appreciable period, perhaps for as long as the completion of a half-hour recording. The latter would appear to be much less efficient. A comprehensive survey of the use of recorded materials in language instruction is provided by Dines (1960), who makes instructive comments on the existing differences and potential similarities between language laboratory instruction and instruction by teaching machines.

Individual programed instruction (auto-instructional methods)

The development of teaching machines and programs for self-instruction requiring the participation of the student has proceeded rapidly during the past few years. Many proponents of teaching machines believe that the auto-instructional concept and its implementation are likely to have very important implications for the conduct of instruction, and may even revolutionize education.

Basically, teaching machines present sequences of programed instruction to the student, requiring his active response at frequent intervals or, in many instances, almost continuously, and providing him with prompt feedback in the form of correction or confirmation for each response. By allowing individuals to proceed at their own rate through a program of materials, they permit students to proceed as rapidly as their abilities permit. The obvious economy of really doing something about individual differences in education (other than merely measuring them) has, in itself, potentially revolutionary significance, not only for the efficiency of instruction, but for the organization of classes and schools.

The literature describing development activities and a certain amount of research on auto-instructional techniques extend back to the mid-1920's following the development of Pressey's first teaching machine. The relatively small stream of papers stemming from Pressey's work expanded enormously following publication of Skinner's 1954 paper in the Harvard Educational Review, and papers based on further research are appearing constantly. Virtually all of this literature up to mid-1960 is either reproduced in full or abstracted in the recent book of readings edited by Lumsdaine and Glaser (1960). The extent of the work summarized in this compilation and in the briefer compendium edited by Galanter (1959), comprising a total of nearly a thousand pages of literature, indicate the emphasis being placed on development in this area at the present time. Papers based on further research and development are appearing so rapidly that it appears likely that a considerably greater number of papers will have been produced by the end of 1962. The work that has been reported deals with the development of mechanical devices and programed materials, with rationale and theory, and with experiments designed either to evaluate the educational effects of programed auto-instructional materials or to investigate specific factors related to their effectiveness. The writers share the opinion that programed instruction (or auto-instructional methods) may represent the most important innovation in education since the advent of the textbook. One reason for this opinion is the fact that the development of programed self-instructional material is closely tied to the use of empirical data obtained from students' responses as a basis for program revision.

## Importance of instructional media

The importance of educational media has two aspects--the role of media vehicles of instruction and their role as objects (or tools) of research.

As a vehicle of instruction. Media are important because of their reproducible characteristics and the fact that resources can be brought to bear that could not readily be otherwise available. Also, because of the fact that they are reproducible and will be used widely, one can afford to spend more money on them in terms of development and research than would otherwise be feasible.

As a research tool and an object of study. Instructional media, because of their reproducible characteristics and the precision with which the stimulus they provide can be displayed (if not always analytically described), have a peculiar advantage, as compared with research on instructional "methods", as tools and objects for research.

## Some important properties of instructional media

Stimulus control. As objects of controllable experimentation that may generate usable research findings, instructional media possess an important characteristic that distinguishes them from instructional methods or procedures. As will be further discussed later, it can be very misleading to reify a specific medium or category of instruments--such as "films", or "television" or "graphic aids"--and to invest it with class properties. Nevertheless, a particular instrument or specific set of instruments, as distinguished from a method or procedure of instruction, constitutes a tangible product or thing. This property of concreteness or thingness has important consequences with respect to constancy, specifiability and *reproducibility* which affect the degree to which experimental manipulation of the properties of instruction can be defined and specified.

For one thing, we can control and specify the experimental stimulus to a degree not often possible when we seek to experiment with properties of a method or procedure of instruction. In the case of a film, for example, we can point to, display and describe very concretely what the experimental stimulus is. The same is true for a self-instructional program designed for use in a teaching machine, or for a tape recording to be used in language instruction. The extent to which we can reproduce the stimulus material in constant fashion is, despite possibly quite important variations in the conditions and context of its use, still far greater than the reproducibility of the stimulus conditions afforded by so-called "methods" or "procedures" of instruction used by a classroom teacher. It is obvious that such constancy, if not indeed *sine qua non* for successful research on instruction, at least greatly facilitates such research.

Dealing with the properties of a concrete product can, furthermore, give us much greater control of the application of research findings than is true for research findings that pertain to abstractions such as a "procedure" or "method". *Products*--things--can be made the subject of *specification*, in a way that hardly applies to "methods"; specifications can be checked for adherence, and specified products in the form of devices and materials have, in turn, an inherent constancy and predictability which can scarcely be expected in the fluctuations that are sure to occur in using method or procedure.

Another important consequence of the reproducibility of materials and devices, related to the better specification and control of their properties as instructional vehicles, is that their production involves fewer people than does the implementation of procedures or methods by a multitude of individual instructors. These individuals are likely to comprise groups that are less dispersed (geographically) and hence more accessible to influence than are individual instructors. One film produced by a relatively small group of individuals is used hundreds of times; there is one act of production, on the part of this small group, which needs to be influenced--i.e., to implement a demonstrably superior technique of presentation--on one occasion, instead of having to insure the utilization of this technique by hundreds of instructors over an indefinite period of time. Similarly, self-instructional programs for a teaching machine, being capable of indefinite reproduction, are relatively easy to control because they are the product of a relatively small group of individuals who can be given specialized training in such skills as the sequencing of materials, to a degree that would seem economically and sociologically quite unfeasible to instill into the multitude of individual teachers of the subject covered by the program. To the extent that these factors conduce to the more direct opportunity to effect utilization of research findings, it can also be argued that higher priority can be accorded, in terms of likely payoff, to research that bears on the controllable properties of instructional media. In fact, it might be held that the easiest way to get principles of effective instruction implemented is to translate them into guidelines or specifications for materials and devices, and to work closely with those who are engaged in producing these tangible instruments of instruction.

The control of student response by instructional media. So far we have discussed stimulus properties and the relative stimulus constancy which inheres in the concretization of instructional properties in the form of tangible materials or instruments. What about the response side as a factor in learning from instruction? Disregarding at least temporarily such questions of interpretation as the exact role of reinforcement in learning and the precise mechanics by which unitary response acts are integrated into complex sustained performances, perhaps it will be agreed that learning involves the development of new response capabilities that were not elicitable, prior to learning, by a given stimulus configuration or situation. In a broad sense, this essential notion surely applies whether we are speaking of response "shaping" or are concerned only with bringing an already well-defined response under the control of a new set of stimuli. Perhaps we may also agree, regardless of differences in specifics of interpretation, that for the acquisition of new stimulus-response contingencies a sine qua non is some form of occurrence of the to-be-learned responses during the course of learning or acquisition--regardless of whether these responses are implicit or overt, simple or complex, unitary or integrated, symbolic or otherwise, mediating or terminal. (cf., Kendler, 1961)

In this sense the responses made by the learner comprise a crucial class of independent variables that influence the outcomes of learning experiences. In another sense, however, student response is never an independent variable, but only a mediating one. This is because we never can manipulate responses *as such*; we can manipulate only the stimulus conditions that foster or permit the occurrence of responses. In this sense, control over the process of learning reduces simply and wholly to the control of stimuli, insofar as it can in fact be manipulated at all, through the control of factors external to the learner himself.

But this by no means implies that the response factors in learning are to be ignored. The authors would contend that we can and should attempt to introduce manipulation of response factors (though the theories and procedures which we bring to bear in so doing must work through the manipulation of stimulus or environmental variables). Indeed, the authors believe that the most productive orientation for research which seeks to improve the effectiveness of instructional media is to look for ways in which appropriate responses of the student, overt and implicit, can be controlled effectively by the stimulus conditions afforded here during the process of learning.

It should be clear that the control of student response can reside in considerable part in the instructional materials--particularly in the way in which they *call for* or *prompt* appropriate response elements of the to-be-learned behavior. To an appreciable extent, however, the occurrence of appropriate student response may depend on the physical opportunities provided for the performance of overt response--by writing pressing keys or other overt manipulation. One reason for the special interest in the instructional potentialities of teaching machines, as opposed to usual forms of films and other audio-visual media, is that the opportunity, and indeed, requirement for overt response is inherent in the instructional situation, and thus forces us to pay attention to the control of student response.

---

# Chapter 2

# Educational Media as Objects of Research

## Education, instruction and media

Since this book deals with "educational media", we should perhaps be more specific about which media we are going to discuss. The term "instructional media" may be substituted for "educational media", since our primary concern is with the effects of using educational media in attaining definable instructional goals. "Educational media" refers to media for presenting or administering an educational "program" or lesson in a sequenced, pre-programmed form. Specifically included are films, televised instruction and "programmed instruction" (the last means programmed in the broader sense as well as in the narrower meaning ascribed to programmed "self-teaching" or "auto-instructional" materials used in a "teaching machine" or "programmed text"). In general, "educational media" is used herein more or less interchangeably with "instructional media", though sometimes perhaps with a slightly more specific connotation indicated in the word "instruction" as compared with "education". By instruction we mean any form of teaching intended to achieve a definite specifiable educational outcome. Instruction thus has a connotation somewhat less broad than does "education" because of the emphasis on quantitatively specifiable objectives or outcomes being sought (which might not be true of some of the more diffuse aims of education).

# Chapter 2

# Educational Media as Objects of Research

## Education, instruction and media

Since this book deals with "educational media", we should perhaps be more specific about which media we are going to discuss. The term "instructional media" may be substituted for "educational media", since our primary concern is with the effects of using educational media in attaining definable instructional goals. "Educational media" refers to media for presenting or administering an education "program" or lesson in a sequenced, pre-programmed form. Specifically included are films, televised instruction and "programmed instruction" (the last means programmed in the broader sense as well as in the narrower meaning ascribed to programmed "self-teaching" or "auto-instructional" materials used in a "teaching machine" or "programmed text"). In general, "educational media" is used herein more or less interchangeably with "instructional media", though sometimes perhaps with a slightly more specific connotation indicated in the word "instruction" as compared with "education". By instruction we mean any form of teaching intended to achieve a definite specifiable educational outcome. Instruction thus has a connotation somewhat less broad than does "education" because of the emphasis on quantitatively specifiable objectives or outcomes being sought (which might not be true of some of the more diffuse aims of education).

The word "instruction" is here used as a generic term, referring to any specifiable means of controlling or manipulating a sequence of events to produce definable modifications of behavior through learning. We are dealing only with those outcomes of learning which can be specified in sufficiently explicit terms to permit their measurement. As noted in Chapter 1, these outcomes may include changes in attitudes, interests, motivations, beliefs or opinions so long as these are defined in terms which identify specific, observable behaviors agreed to be manifestations or indications of these outcomes. Use of the word "outcomes" rather than "objectives" does not, however, preclude considerations of *any* outcome--whether anticipated, desired or otherwise--even though our main concern is with the attainment of objectives which are chosen in advance to represent the purpose of the instruction (cf. Chapter 1, pp. 17-18).

As used in this book, "*instructional media*" refers either to a particular class of instructional resources (e.g., films), or to a group or subset of such a class, or to media collectively. An instructional "*instrument*" or "*program*" generally refers to a particular member or instance of this class, e.g., a particular film, self-instructional program, televised course, etc. The terms "instrument" and "media" can refer either to the instructional materials prepared for teaching a particular subject matter, or to the combination of equipment and associated materials--e.g., a film and projector, or a teaching machine and the program materials presented by it.

Instructional resources which can aid in the learning process include, in addition to the teacher, a wide variety of tools, devices, materials and other "aids". Broadly speaking, these either present, or help to present, information to the learner, or afford opportunities for him to make responses which help him to learn or indicate what he has learned. The "audio-visual" media of instruction of course comprise an important and widely used class of such instructional resources, generally employed to present information. Other instructional tools include instruments which permit the student to make relevant responses (e.g., workbooks, self-scoring test forms and other manipulanda) which can aid the learning process. The use of "programmed instruction" materials, particularly teaching machines or "auto-instructional" devices, serves both an information-presenting function and also affords a means for the student to make appropriate responses to the material presented.

The term programed instruction means, generically, any pre-sequenced instruction given by instructional media to groups of individuals. More specifically, depending on context, this term may refer more narrowly to forms of programed instruction (or "programed learning") which use teaching machines or programed-text material to present individual instruction to a single student who is learning at his own rate, with the additional characteristics of requiring frequent response by the student and giving him immediate feedback (confirmation or correction) for each response he makes. Note that one of these characteristics can be present without others; in particular, frequent student response and immediate feedback may be employed in group instructional media by films or television.

Programed individual instruction further implies an individually determined rate of presentation and response and permits the student to proceed for an appreciable time without the intervention of a teacher. This latter characteristic is also true of many forms of audio-visual media, e.g., sound films, in which the student is supposed to learn from the audio-visual presentation without the continuous intervention of the teacher, just as he is expected to do when pursuing independent study of a textbook or reference source. Some other visual and audio-visual "aids" by contrast, are used by the teacher as adjuncts to his other activities, rather than as independent devices of instruction. (All instructional instruments and the media they present can properly be considered "aids" to learning. However, not all need be considered aids to the teacher, since some, e.g., self-instructional programs, or in a sense, textbooks, are clearly capable of instructing without the teacher's direct intervention.)

### Educational Media as sequenced and reproducible presentation

The educational media with which this book deals are primarily those presented in some predetermined sequence. The sequence is determined either by the structure of the instructional instrument, by well defined procedures and constraints in the way they are used, or by the mechanical arrangements used to present them. The instruments thus generate events whose occurence is more or less predictable in terms of sequence or pacing. The sequencing and pacing need not be fixed, but they do depend on clearly specifiable factors in the learning situation. A specifiable factor does not mean merely the activities of an instructor, nor the decisions of a student in whose hands materials are placed. This restriction does not mean that media presentation may not require the activity of a teacher or assistant (e.g., film projectionist); it does mean that the student is expected to learn something from a previously prepared sequence of instruction even though the teacher is not present or active at the moment. Instructional media, as the term is used here, provide acts of instruction, not merely the wherewithal for such acts.

The events generated by instructional media also tend to be relatively reproducible events in that they may be repeated (as in the case of a film) or be duplicated (as in the case of instructional television) at a number of locations simultaneously. In concrete terms, the instructional instruments dealt with here include motion picture films, tape recordings, filmstrips, and self-instructional programs for presentation by a teaching machine or some related device. Instructional programs or instruments do not, in general, include charts, diagrams, models or isolated photographs or drawings, as such, except when these represent elements in a predictable pattern of instruction governed by a relatively well defined procedural sequence. Such isolated elements can be regarded only as tools or components for instructional acts or instruments; a complete instrument consists of the aid plus the activities of the teacher. This restriction is necessary if we are to consider instructional media and instruments as specifiable events which are amenable to scientific research, which can deal only with reproducible events.

As the senior author has pointed out elsewhere (Lumsdaine, 1960) single visual displays represent only an isolated and incomplete fragment of a total instructional stimulus pattern. Thus, except when they are considered only as a component of other specified instruction, it is difficult to control or predict their practical effects on instruction. When they are considered alone, it is difficult to partial out their contribution from other unspecified factors in the instructional situation. Perhaps this is why little significant experimental research has been done on non-sequenced, isolated visual aids.

In terms of the criteria set forth above textbooks represent a borderline case. The usual textbook does not control the behavior of the learner in a way which makes it highly predictable as a vehicle of instruction or amenable to experimental research. It does not in itself generate a describable and predictable process of learner behavior, and this may be the reason why there has been very little productive experimental research on the textbook. In any case, the textbook and its cousin, the training manual or training syllabus, are not treated explicitly here as a primary object of research.

## Major Classes of Instructional Media

### Instructional films and related media

The terms "teaching films," "educational films," "training films" (the military and industrial term) all indicate presentation of materials by motion-picture film as used for instructional [illegible] Most instructional films are currently produced as 16 mm. sound films, in black and white or in color. They present an enormous variety of instructional materials, including film-recorded lectures, surrogate field trip visits (the travelogue and other films which aim primarily to "bring the world to the classroom"), dramatic or theatrical presentations of literary or historical events, and dramatized forms of exposition which depict the teaching of students within the film by a teacher or parent. In addition, two quite different kinds of "demonstrational" films should be distinguished; first, demonstrations of *phenomena* depicting events or characteristics of the animate or inanimate world and, second, *procedural* demonstrations designed to teach the performance of a specific sequence of acts.

*Instructional objectives of films and related media.* Aside from some instances in the latter category, sometimes deprecatingly referred to as "nuts and bolts" training films, the instructional objectives for which films were presumably designed unfortunately tend to be stated imprecisely, if at all; when not characterized merely as "telling a story about" a subject matter, or "presenting" the "content" of a lesson (without specification as to what precisely is to be learned), the objectives are often described as those of creating a general familiarity, understanding, background, or "knowledge about" the subject treated. Because of this vagueness of objective, it is often difficult to define the appropriate criteria for evaluating instructional films. That the same often applies to other, non-media forms of instruction is perhaps less often obtrusive because of the fact that non-media instruction is less often made the subject for experimental evaluation.

*Film production.* Most instructional sound films are now produced either with synchronized sound recorded concurrently or with narration and other sound added separately. Their high cost when commercially produced can sometimes be very greatly reduced by local, short-cut production techniques such as those described by Greenhill (1955).

Often the content and sequencing of instructional films are governed in part by practices and traditions borrowed from the field of theatrical motion pictures as much as by considerations of pedagogy.

_Films and kinescopes or video tape recordings._ The respects in which films are likely to differ instructionally from a kinescope or video tape recording of live televised instruction probably lie more in philosophy and practice of production than they do in the inherent characteristic differences of the media. (Television not prefilmed or recorded by video tape or kinescope is, of course, transitory and not temporally repeatable.) Aside from minor differences in grain or resolution (these differences are very considerable in terms of physical measurement but probably are seldom important from an instructional standpoint), the televised instructional lesson, kinescope recording, and video tape recording differ instructionally from the sound motion picture primarily with respect to screen size and in the potentiality for using color, except in cases where immediacy of presentation is a genuinely important factor. Even this difference is more a matter of tradition and usage than of inherent properties, however, since motion picture film can, by special equipment, be processed and projected within a matter of seconds from the time an event is photographed. For most instructional purposes, we shall therefore consider the inherent instructional properties of film and of recorded television presentations as substantially identical. Both consist, in fact, of a series of moving or sequenced still pictures accompanied by recorded sound, and in many instances only an expert can detect the difference between pre-filmed or taped and live television presentations, or between films photographed directly on a photographic emulsion and those in which the original action was picked up by a television camera.

*Differences in production practices*. At the present time, however, films and televised instruction often differ considerably because of tradition-engendered factors related to inherent or adventitious constraints in the budgets and time schedules for their production, and also because of conventionalpractices such as that of filming television-linked films "straight through" (in contrast to standard practices in producing films by shooting one scene at a time, and later putting the resulting bits and pieces together by a careful job of editing in the cutting room.) This difference in tradition appears to be weakening, however, with the increasing use of pre-filming for both instructional as well as entertainment television, and of the so-called "straight through" or "television-style" techniques. (The feasibility of the latter is somewhat enhanced by the so-called "telecam" technique in which the same image is picked up concurrently on film and over a video circuit, enabling on-the-spot direction and editorial decisions to be made, while at the same time preserving a high quality film original for subsequent reproduction.)

Leaving aside these questions of production technique, most of the variation in cost as well as instructional properties results from what is placed in front of the camera. It may also be noted (e.g., in the subsequently described Lumsdaine-Sulzer-Kopstein study: 1951, 1961) that the important characteristic of reproducible, carefully planned instruction, as represented by film and its TV cousins, can be instructionally valuable even where the visual material is largely of a static sort.

Similar instruction in this case can often be provided by a sound-accompanied series of still pictures, requiring simpler equipment and less expensive materials in some instances.

Semi-motion presentations. Some forms of devices for sequenced still-picture projection are not only capable of regulating the timing of successive frames or scenes by automatic signals from the accompanying sound recording, but also are capable of changing frames rapidly enough that simulated motion, pop-on effects, and other characteristics of motion-picture presentation are closely simulated. When one considers also the actual absence of essential motion, and frequent lack of need for it, in film sequences running many feet in length of motion-picture film, it is evident that the distinction between "film" and "filmstrip" media is a hazy and shifting one, clear only in the extremes and even then better specified in terms of specific stimulus properties of motion and transition than by characterization in terms of "media" properties.

For group-presented pre-paced (as well as pre-sequenced) presentation, it is true, however, that presentation by motion-picture film projector has maximum flexibility in the sense that it permits any combination of still and motion presentation. On the other hand, the motion picture is sometimes at an important cost disadvantage where motion is not really utilized effectively for considerable periods of the instruction. In addition, flexibility is lost since, with most existing forms of motion picture projection equipment, at least the pacing or timing of the instructional material cannot be guided by the reactions of an individual learner. This has thus far not been considered a primary disadvantage because of the fact that film presentation has generally been used as a group medium. However, with the advent of teaching machines and the concepts associated with them, increasing use will be made of individually presented films having flexible start-stop characteristics.

_Film utilization in education_. It is difficult to obtain firm data on the prevalence of film usage; no exact counts are readily available on the total number of educational films in current use in our schools and colleges, nor on the total number of showings or pupil hours of film viewing presented per year. The total number of films available, however, is very large; the most comprehensive single catalogue, Wilson's lists over 10,000 films in the 1954-58 edition and subsequent supplements. Teachers have increasingly been trained and encouraged to use films, as well as the simpler visual aids to instruction; some school systems require teachers to have had at least a one unit laboratory course in the operation of film projectors and related equipment.

Characteristic guidelines for the use of films have stressed the role of films as _aids_ to teaching, rather than as self-contained instruments of instruction. Teachers are encouraged to preview films, to provide introduction to the films for students, to conduct followup discussions, and less commonly, to intersperse question and answer sessions or other class participation exercises between sections of a film. Although these practices are unquestionably useful in many instances, the perspective of regarding the film merely as an adjunct to other instruction is by no means a necessary one, and there has been an increasing tendency to construct films which are intended to stand on their own feet, if necessary, as independent instructional media. A well-known example is the Harvey White physics series, which was initially filmed on the basis of a series of televised lessons. One suspects that the emphasis on using films as aids is one which is slowly changing as better and more comprehensive films become available, and that it has been, in the past, rooted partly in the knowledge that many films are ill-equipped to stand on their own feet as independent acts of instruction, and partly in apprehension lest the film displace the teacher. It is interesting to note that, in some instances, recent uses of films are perceived as carrying the main burden of instruction with the classroom teacher aiding the film presentation, rather than the film being used as an aid to a sequence of instruction organized by the teacher. An advantage provided by the "self-sufficient" televised or filmed lesson (shared with self-instructional programed materials) is that it makes it possible to offer courses of instruction in a wide variety of subjects in small schools where qualified teachers are not available for all of these subjects.

### Further aspects of television

Instructional television has tended even more than films to assume the primary role in conveying instruction; even where a teacher is present, her role is often a supporting rather than a supported one. Where complete half-hour or hour-long lesson periods are transmitted by television, the classroom teacher in some instances provides follow-up, or preparatory activities; in other instances, the televised instruction in a particular subject is the only instruction that the students get. Whatever disadvantages may be raised for this latter pattern, it represents, evidently, an increasing trend. However, closed circuit television has also been utilized as an aid to the instructor--for example, in the use of the overhead television camera as a tool at the disposal of the lecturer to replace slide projections and other forms of projected aids, or to transmit demonstrations, e.g., in science and medicine, including surgical operations, so that more students can get a good view of the demonstration. However, the primary role has been that of televising a complete lesson, and the increasing tendency to record these lessons--and the obvious wastefulness of not doing so when extensive preparation and resources have gone into their production--makes it further apparent that, for many purposes, televised instruction and equivalent instructional films are identical media. Certainly, research on factors that make for effective presentation of information on a film screen should, in most particulars, apply to equivalent instruction projected on a television screen.

Broadcast and closed-circuit TV instruction. In addition to the "educational" programs transmitted over television for public information or general enlightenment, many class hours of instruction are being transmitted over educational stations in the United States. Approximately [illegible] stations are licensed at the present time as educational stations on UHF or VHF channels, and many of these devote considerable time, particularly during the school day, to televised lessons for use in school classrooms or as extension credit courses. For example, WQED in Pittsburgh and its UHF affiliate, WQEX, together currently transmit over eight hours a day, including lessons in subjects ranging from shorthand to physics and providing basic course content for approximately ten courses in addition to supplemental programs for numerous other subjects. Extensive coverage is also provided by an increasing number of closed-circuit installations operated by school systems (e.g., Hagerstown, Maryland, and Anaheim, California). The advent of an airborne television facility designed to blanket a wide geographic region with televised courses further dramatizes the increasing utilization of transmitted instruction and emphasizes the importance of research on presentation characteristics.

*TV as an aid*. As noted above, the use of television is not limited to transmitting an entire lesson or series of lessons over a closed-circuit or open-circuit (broadcast) channel; it may also be employed as an aid to an instructor in a classroom. Bretz, at the University of California, Los Angeles, has developed a portable installation which can be quickly set up in most classrooms and which employs an overhead camera at the instructor's podium to substitute for a variety of other projection facilities and to eliminate the need for special transparencies. Textbook illustrations, charts, drawings, and other visual material which the instructor wishes to use in a lecture are simply placed on the desk under the camera and are projected on one or more monitor screens in the classroom. The use of a zoom lens on the overhead camera permits adaptation to materials of different sizes by increased magnification for showing detailed portions of a display. (See Wade and Bretz, 1960.)

## Audio Instruction

Unquestionably the major current use of audio instruction in most schools in the United States is for teaching foreign languages. Although a number of recorded lectures have been made available to schools, their use and that of radio-presented instruction has never caught on in the United States as the use of films and TV-presented instruction has. Thus far, less attention appears to have been paid to making well-designed instructional recordings available than to the installation of language laboratory equipment, the assumption evidently being that the teacher would prepare materials locally for use with the equipment.

Hundreds of such language laboratory installations have been made, generally including a magnetic recording system employing tapes or, occasionally, discs. Students listen to recorded material in the foreign language and then attempt to reproduce it. A number of the systems allow recording of the student's effort so that he may compare it with the master recording he has imitated. The effectiveness of such instruction depends not only on the availability of recorded materials, but also on the characteristics of the equipment. In some instances, this permits the student to proceed at his own pace, using an individual recording. In other instances, the group is lock-stepped to a common master recording. In some systems, the student is able to play back a comparison of his own vocalization efforts *immediately* in comparison with the model; in other systems, playback and comparison has to be delayed for an appreciable period, perhaps for as long as the completion of a half-hour recording. The latter would appear to be much less efficient. A comprehensive survey of the use of recorded materials in language instruction is provided by Haas (1960), who makes instructive comments on the existing differences and potential similarities between language laboratory instruction and instruction by teaching machines.

Individual programmed instruction (auto-instructional methods)

The development of teaching machines and programs for self-instruction requiring the participation of the student has proceeded rapidly during the past few years. Many proponents of teaching machines believe that the auto-instructional concept and its implementation are likely to have very important implications for the conduct of instruction, and may even revolutionize education.

Basically, teaching machines present sequences of programmed instruction to the student, requiring his active response at frequent intervals or, in many instances, almost continuously, and providing him with prompt feedback in the form of correction or confirmation of each response. By allowing individuals to proceed at their own rate through a program of materials, they permit students to proceed as rapidly as their abilities permit. The obvious economy of really doing something about individual differences in education (other than merely measuring them) has, in itself, potentially revolutionary significance, not only for the efficiency of instruction, but for the organization of classes and schools.

The literature describing development activities and a certain amount of research on auto-instructional techniques extend back to the mid-1920's following the development of Pressey's first teaching machine. The relatively small stream of papers stemming from Pressey's work expanded enormously following publication of Skinner's 1954 paper in the Harvard Educational Review, and papers based on further research are appearing constantly. Virtually all of this literature up to mid-1960 is either reproduced in full or abstracted in the recent book of readings edited by Lumsdaine and Glaser (1960). The extent of the work summarized in this compilation and in the briefer compendium edited by Galanter (1959), comprising a total of nearly a thousand pages of literature, indicate the emphasis being placed on development in this area at the present time. Papers based on further research and development are appearing so rapidly that it appears likely that a considerably greater number of papers will have been produced by the end of 1962. The work that has been reported deals with the development of mechanical devices and programed materials, with rationale and theory, and with experiments designed either to evaluate the educational effects of programed auto-instructional materials or to investigate specific factors related to their effectiveness. The writers share the opinion that programed instruction (or auto-instructional methods) may represent the most important innovation in education since the advent of the textbook. One reason for this opinion is the fact that the development of programed self-instructional material is closely tied to the use of empirical data obtained from students' responses as a basis for program revision.

Importance of instructional media

The importance of educational media has two aspects--the role of media vehicles of instruction and their role as objects (or tools) of research.

As a vehicle of instruction. Media are important because of their reproducible characteristics and the fact that resources can be brought to bear that could not readily be otherwise available. Also, because of the fact that they are reproducible and will be used widely, one can afford to spend more money on them in terms of development and research than would otherwise be feasible.

As a research tool and an object of study. Instructional media, because of their reproducible characteristics and the precision with which the stimulus they provide can be displayed (if not always analytically described), have a peculiar advantage, as compared with research on instructional "methods", as tools and objects for research.

Some important properties of instructional media

Stimulus control. As objects of controllable experimentation that may generate usable research findings, instructional media possess an important characteristic that distinguishes them from instructional methods or procedures. As will be further discussed later, it can be very misleading to reify a specific medium or category of instruments--such as "films", or "television" or "graphic aids"--and to invest it with class properties. Nevertheless, a particular instrument or specific set of instruments, as distinguished from a method or procedure of instruction, constitutes a tangible product or thing. This property of concreteness or thingness has important consequences with respect to constancy, specifiability and reproducibility which affect the degree to which experimental manipulation of the properties of instruction can be defined and specified.

For one thing, we can control and specify the experimental stimulus to a degree not often possible when we seek to experiment with properties of a method or procedure of instruction. In the case of a film, for example, we can point to, display and describe very concretely what the experimental stimulus is. The same is true for a self-instructional program designed for use in a teaching machine, or for a tape recording to be used in language instruction. The extent to which we can reproduce the stimulus material in constant fashion is, despite possibly quite important variations in the conditions and context of its use, still far greater than the reproducibility of the stimulus conditions afforded by so-called "methods" or "procedures" of instruction used by a classroom teacher. It is obvious that such constancy, if not indeed sine qua non for successful research on instruction, at least greatly facilitates such research.

Dealing with the properties of a concrete product can, furthermore, give us much greater control of the application of research findings than is true for research findings that pertain to abstractions such as a "procedure" or "method". Products--things--can be made the subject of specification, in a way that hardly applies to "methods"; specifications can be checked for adherence, and specified products in the form of devices and materials have, in turn, an inherent constancy and predictability which can scarcely be expected in the fluctuations that are sure to occur in using method or procedure.

Another important consequence of the reproducibility of materials and devices, related to the better specification and control of their properties as instructional vehicles, is that their production involves fewer people than does the implementation of procedures or methods by a multitude of individual instructors. These individuals are likely to comprise groups that are less dispersed (geographically) and hence more accessible to influence than are individual instructors. One film produced by a relatively small group of individuals is used hundreds of times; there is one act of production, on the part of this small group, which needs to be influenced--i.e., to implement a demonstrably superior technique of presentation--on one occasion, instead of having to insure the utilization of this technique by hundreds of instructors over an indefinite period of time. Similarly, self-instructional programs for a teaching machine, being capable of indefinite reproduction, are relatively easy to control because they are the product of a relatively small group of individuals who can be given specialized training in such skills as the sequencing of materials, to a degree that would seem economically and sociologically quite unfeasible to instill into the multitude of individual teachers of the subject covered by the program. To the extent that these factors conduce to the more direct opportunity to effect utilization of research findings, it can also be argued that higher priority can be accorded, in terms of likely payoff, to research that bears on the controllable properties of instructional media. In fact, it might be held that the easiest way to get principles of effective instruction implemented is to translate them into guidelines or specifications for materials and devices, and to work closely with those who are engaged in producing these tangible instruments of instruction.

The control of student response by instructional media. So far we have discussed stimulus properties and the relative stimulus constancy which inheres in the concretization of instructional properties in the form of tangible materials or instruments. What about the response side as a factor in learning from instruction? Disregarding at least temporarily such questions of interpretation as the exact role of reinforcement in learning and the precise mechanics by which unitary response acts are integrated into complex sustained performances, perhaps it will be agreed that learning involves the development of new response capabilities that were not elicitable, prior to learning, by a given stimulus configuration or situation. In a broad sense, this essential notion surely applies whether we are speaking of response "shaping" or are concerned only with bringing an already well-defined response under the control of a new set of stimuli. Perhaps we may also agree, regardless of differences in specifics of interpretation, that for the acquisition of new stimulus-response contingencies a sine qua non is some form of occurrence of the to-be-learned responses during the course of learning or acquisition--regardless of whether these responses are implicit or overt, simple or complex, unitary or integrated, symbolic or otherwise, mediating or terminal. (cf., Kendler, 1961)

In this sense the responses made by the learner comprise a crucial class of independent variables that influence the outcomes of learning experiences. In another sense, however, student response is never an independent variable, but only a mediating one. This is because we never can manipulate responses as such; we can manipulate only the stimulus conditions that foster or permit the occurrence of responses. In this sense, control over the process of learning reduces simply and wholly to the control of stimuli, insofar as it can in fact be manipulated at all, through the control of factors external to the learner himself.

But this by no means implies that the response factors in learning are to be ignored. The authors would contend that we can and should attempt to introduce manipulation of response factors (though the theories and procedures which we bring to bear in so doing must work through the manipulation of stimulus or environmental variables.) Indeed, the authors believe that the most productive orientation for research which seeks to improve the effectiveness of instructional media is to look for ways in which appropriate responses of the student, overt and implicit, can be controlled effectively by the stimulus conditions afforded here during the process of learning.

It should be clear that the control of student response can reside in considerable part in the instructional materials--particularly in the way in which they call for or prompt appropriate response elements of the to-be-learned behavior. To an appreciable extent, however, the occurrence of appropriate student response may depend on the physical opportunities provided for the performance of overt response--by writing, pressing keys or other overt manipulation. One reason for the special interest in the instructional potentialities of teaching machines, as opposed to usual forms of films and other audio-visual media, is that the opportunity, and indeed, requirement for overt response is inherent in the instructional situation, and thus forces us to pay attention to, the control of student response.

# Chapter 3

# The Nature of Experiments

As stated at the outset, the concern of the volume is with an analysis of methods used and results obtained in experiments on the development of knowledge, skills and attitudes, through educational communication media. The word "experimental" figures in the title, and the word "experiments" figures further in the heading of this chapter. The first thing we need to do is to clarify somewhat further what we mean by an experiment.

### The crucial nature of elementary considerations

In this chapter and the next few chapters we are going to try to give a clearer notion of what an experiment is, discuss some of the basic characteristics of experiments, and consider some of the methodological problems that arise in conducting them. In doing so we shall be dealing at a rather elementary (yet nonetheless fundamental) level with a number of terms and concepts which the reader may already be familiar with. Parts of our discussion may seem so elementary as hardly to be worthy of notice for experienced investigators. We should like to stress our conviction that this is not the case. On the contrary, in our view, an attempt to return to fundamental considerations and deal with them in elementary fashion is absolutely essential to dealing clearly with advanced methodological problems--such as those that enter into some of the educational-media research of the kinds discussed in Chapters 13 and 14. It is our feeling that many of the shortcomings of research *which has been done in the instructional-media field stem rather directly from the failure to keep clearly in mind some of the "kindergarten" considerations presented here.*

The purpose of this chapter and those that immediately follow it is to delineate and illustrate some very basic aspects of experiments--aspects essential to clear understanding of specific features of experiments on instructional media that will be dealt with later. An overview of these aspects is essential to the understanding and evaluating of specific experiments such as those whose results are summarized in Chapters 13 and 14.

*Successful and unsuccessful experiments*. We are going to try to arrive at this meaning partly by a process of successive approximation to achieve a clearer and clearer idea of what constitutes an experiment. In doing this we will necessarily have to consider what makes for successful and unsuccessful experiments, and will gain some notion of how difficult it is to formulate and carry out a really useful and successful experiment. We will also need to see how easy it is to do something that *looks like* a good experiment but that is fated, in advance, to be unsuccessful for lack of sophistication about how to do it. This is not intended to sound forbidding. On the contrary, the *essential* things about an experiment are rather simple, at least in principle.

## Rational and empirical inquiry

First let us try to delineate the field of experimental research a bit more sharply by making some distinctions. A dictionary definition of research is "careful or critical examination in seeking facts or principles." Such inquiry may, first of all, be characterized as either "rational" or "empirical". As applied to educational media, rational inquiry would involve reflection, formulation of hypotheses, theorizing, or other "armchair" or speculative forms of inquiry. Related to such inquiry is the examination of historical records and similar data, or the critical analysis of previous opinion and theorization. Such rational inquiry may often make an important or essential contribution to empirical research, but is not in itself empirical in that it does not obtain new data through systematic observation. Empirical inquiry, by contrast, involves the obtaining and analysis of new data--in our case, data about the presumed effects of instructional media. Empirical data might, however, include a great deal of information not dealt with herein--such as the data obtained by "survey" methods, including statistics on the distribution and use of media, for example, or opinions about educational objectives and ways to attain them, or test results indicating existing levels of achievement among school populations.

Inductive inference. The drawing of conclusions about cause-effect relationships from empirical data involves the process of inductive inference. It is our position that only in this way can hypothesized causal relationships be demonstrated and, hence, can our knowledge of the effects of instructional media be increased. It is for this reason that we devote the present book to the clarification of the experimental method as applied to instructional media. As R. A. Fisher (Design, Sec. 4, p. 7) has put it:

> Inductive inference is the only process known to us by which essentially new knowledge comes into the world. To make clear the authentic conditions of its validity is the kind of contribution to the intellectual development of mankind which we should expect experimental science would ultimately supply.

Fisher (Design, Sec. 2, P. 4) also has well stated the necessary answer to those who feel that such inference must lack rigor because it is necessarily uncertain:

> We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression. In the theory of probability, as developed in its application to games of chance, we have the classic example proving this possibility. If the gamblers apparatus are really true or unbiased, the probabilities of the different possible events, or combinations of events, can be inferred by a rigorous deductive argument, although the outcome of any particular game is recognised to be uncertain. The mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference.

The problems of induction and generalization are of course crucial in the interpretation of experimental data and will recur at several later points in our discussion. (See especially the section on generalization [including Campbell and Stanley quote] in Chapter 11.

Experiments.

Experimental research--the special kind of empirical research with which we are dealing--has two further characteristics, mentioned previously, which help to define it. First, it seeks observations which attempt to reveal effects of causal factors (these being, in our case, instructional media). And second, in doing so, the experimenter manipulates the causal factors (here, the use of media)--by administering, withholding, changing what is presented according to a plan designed to reveal these effects.

A set of empirical observations is not considered an experiment unless it has both of these characteristics. Merely manipulating the content of a film or "trying out" an instructional technique--without a way of identifying the effects of this on the learner--may result in useful innovating, but does not constitute an experiment as the term is used here. Likewise, attempts to infer the effects of some instructional instrument by observing what happens after use of the instrument but without manipulating (i.e., arbitrarily introducing or withholding it) will not be considered as experiments. The reason for this has already been mentioned briefly in Chapter 1, and will be discussed in more detail presently; basically, the argument is that cause-effect relationships between media and educational effects ordinarily cannot be isolated dependably without deliberate manipulation of the causal agent.

Controlled observation

In talking about experiments, we are dealing with situations that are set up to provide controlled conditions for observation. These conditions are designed to provide a sounder basis for drawing conclusions than is available in the normal course of casual, unplanned or "naturalistic" observation. In an experiment we attempt to arrange conditions in such a way that cause-effect relationships can be unambiguously revealed. For the time being we will accordingly characterize an experiment (or "controlled experiment") on educational media simply as a systematically planned study in which the arbitrary administration of one or more instructional programs under controlled conditions is used, in conjunction with subsequent observation, so as to be capable of yielding a relatively unambiguous assessment of the effects that can be attributed to the program, in the sense that the program can be considered the cause of the observed effects.

The characterization provided by Fisher (Design, Sec. 4 p. 8 ) is pposite here:

> Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge; that is, they are systematically related to the body of knowledge already acquired, and the results are deliberately observed, and put on record accurately.

Reproducibility. An experiment also can be characterized as a trial method or procedure, applied to a particular problem, under reproducible conditions. The method is one that can be re-employed or repeated for purposes of checking results, or in order to develop variations of the experimental interaction. An essential thing about science is that it yields reproducible findings. If one person measures the coefficient of expansion of copper on Tuesday and finds that it is .0002 and another person measures it in a different room on Thursday and finds it to be .0005, we would say that the experiment is not reproducible. Similarly, in an educational experiment, unless we can specify the condition in such a way that the experiment is inherently reproducible, we are defeated before we start. It is therefore important to see what kinds of instructional-media experiments inherently carry with them the seeds of reproducibility and what kinds of features tend to defeat this requirement.

"Controlled experiment." The control of the conditions of observation characteristic of experiments sometimes leads to their being called "controlled experiments", to distinguish them from more informal tryout experiences in which causal factors are not well enough specified to insure reproducibility, or in which extraneous causal factors are not sufficiently controlled to permit an unambiguous attributing of an observed effect to a particular cause. The adjective "controlled" in "controlled experiment" is most often encountered in the areas of inquiry where true experiments are rare--e.g., in sociology or economics. For the most part, the term "experiment" used herein refers to a controlled experiment, the modifier being considered redundant.

Cause and effect

We shall not debate the extent to which cause and effect are epistemologically ultimate concepts. We believe we can--at a practical level that encompasses most scientific experimentation as well as the experiences of everyday life--deal with cause and effect in a sense that we all understand well enough for our present purposes. Basically, what we mean is: If you do such and such, thus and so results. As applied to instructional media, what we mean is: When an individual participates in a given instructional situation, it has some observable effect on him. By effect we mean that, at the conclusion, he is different from the way he would have been without the instruction. (Here, as in Chapter , we may make use of the "before-after" or "with-versus-without" concept, to which we shall again return presently.)

The question of how is he "different" from what he would have been may also be asked with respect to a number of specific possible outcomes. As indicated in ChapterI, these outcomes are called the *effects* of the experiment, meaning the learning that results from the instruction. (The reader should be reminded that when we talk about learning, we mean learning in a very general sense; that is, any modification of behavior, due to experience, with the exception of certain obvious exclusions such as the effects of fatigue.)

## The anatomy of experiments: component elements and terms of reference

Before proceeding further in our delineation of the characteristics of instructional-media experiments and the requirements for valid experimentation, it will be helpful to define some terms. The terms to be defined first are purely descriptive ones that identify the "anatomy" of an experiment, without specific reference to its quality or goodness or success; they are terms which provide convenient labels for the elements and operations in any experiment. For this purpose we will give some provisional definitions or characterizations of several terms which will enter frequently throughout our discussion of experiments. These are working definitions and are not exact ones in all cases. They can be regarded as approximate characterizations of terms for which, in some cases, more exact definitions are also summarized in the glossary at the end of the volume.

Some of the terms that we need in discussing experiments may be discovered by considering the purpose of doing experiments.

Why do you do an experiment? Basically, to answer a question about the effects of something--in our case, the effects of an instructional instrument (program) that is an instance of some form of educational media. In doing an experiment, you start with a question. Not just any question, but a question about the effect of doing something on the result that ensues. The "something" that you do represents a causal factor, something that is manipulated. Often this is called a "treatment." As used herein, an experimental treatment means a specified form or instance of the use of instructional media--that is to say, the use of an instructional instrument or instructional program under conditions of controlled administration in an experiment.

Since the purpose of the experiment is sometimes to compare the effect of different variations, or degrees of the treatment--e.g., varied amounts of repetition, or different degrees (or kinds) of coaching--the manipulated factor also is called an "independent variable." A result that is observed is called a "dependent variable." This means an "effect," representing a change in some indication of performance or capability--as observed in the behavior of the subjects (i.e., students or more generally, learners). Amount learned from a film could be a dependent variable, as could attitude change. The reason for this terminology is fairly obvious: the "independent" variable is a causal factor that is independently introduced; the "dependent" variable is that which "depends on" the independent variable.

Intervening between the independent and dependent variables in all educational experiments is, of course, a person, the learner or student. It is to him (and in a sense for him) that the treatment representing the independent variable is applied, and it is his subsequent behavior, in some form, which we need to observe to provide the dependent variable that can indicate the effect of the instructional treatment. We refer to a learner (or learners) who participates in an instructional or other psychological experiment (and whose behavior yields the dependent variable which embodies the effect of the instruction) as the "Subject" (or "Subjects") in the experiment. "Subject" and "Subjects" are sometimes abbreviated "S" or "Ss", respectively. Since in educational experiments there is sometimes confusion between "subjects" in this sense and "subject" in the sense of subject-*matter*, the term learner or student (or pupil or trainee) is frequently used herein in place of the word "Subject." (The abbreviations *S* and *Ss* can, if desired, conveniently be read "students" in most reports of experiments on instruction.) The terms "task" or "subject-matter" (also "learning task") are used more or less interchangeably here to indicate *that which is to be learned* by the student as the result of receiving instruction. In addition to the use of the abbreviations *S* and *Ss*, we sometimes use in psychological and educational experiment, the abbreviation *E* to stand for Experimenter or investigator.

## *Independent and dependent variables*

The terms *dependent variables* and *independent variables* do not have fully standardized meanings. Independent variables may include anything which is varied by an investigator. They may be varied either by his introducing or manipulating a treatment experimentally, or by differentiating various groups that differ with respect to some population characteristic (see below).

*Two kinds of independent variables*. An independent variable in an educational study can be either of two general kinds. First, it can be a so-called "*selection variable*," or "*sorting variable*." For example, we might compare the knowledge of arithmetic among bright versus dull, better educated versus less well educated, or Jews versus Catholics.

This kind of independent variable can be used to differentiate or sort out data (in terms of some dependent variable) in relation to pre-existing differences amoung groups of people. Doing so may yield useful knowledge but does not isolate causes and is not an experiment.

A quite different kind of independent variable is, for example, the introduction of a film (or of a given amount or kind of repetition in a film). The difference between these two classes of independent variables represents a basic difference between what is an experiment and what is not an experiment. We thus again note that, in an experiment, a dependent variable is observed "as a function of" (i.e., in consequence of) some difference in an independent variable, where the independent variable is manipulated or arbitrarily introduced by the experimenter, instead of being segregated by being taken as he finds it (i.e., through sorting or selection in terms of some pre-existing characteristic).

Experimental variables include those independent variables which are introduced through manipulation by the experimenter by systematically administering, varying or withholding an experimental treatment on some arbitrary basis in order to assess its effects in relation to the experimental hypothesis he wishes to test. The "arbitrary" basis would be one which involves unbiased or, essentially, random assignment of the treatment to certain learners (rather merely than identifying those who happen to have received a treatment, or segregating on an ex post facto basis those who have a particular characteristic). An experimental variable is always considered as an independent variable, but not vice versa.

The term "experimental variable" is frequently restricted (as distinguished from "control variable" or "procedural-control variable") to those independent variables whose hypothesized effects it is the purpose of an experiment to assess--i.e., to the experimentally manipulated variables that are of intrinsic interest to the hypothesis being tested (see also p.21, Chapter 7). An experimental variable is, in any case, one which can be manipulated by the experimenter, by arbitrarily introducing or withholding it. An experiment will always have one experimental variable; it may have more than one.

Some classes of independent variables

"Independent variables" are, generically speaking, the causal factors that enter into experiments as factors which may influence the dependent variables that are observed. Independent variables can be usefully further classified in several ways.

One classification is in terms of the origin or locus of variation. A useful classification in these terms was introduced by Hovland, Lumsdaine and Sheffield (1949). In the case of experiments on films, they distinguished among population variables, film variables and external variables.

Population variables or learner variables are "selection variables" (or "sorting variables": see further discussion in Chapter 7) reflecting previously-determined characteristics of the students (e.g., intelligence, age, and previous -- pre-experimental -- knowledge of the subject matter). They thus are not capable of experimental manipulation within an experiment, though their influence may be studied by sorting the effects of an experimental variable in terms of sub-categories--e.g., effects of a film on bright and dull students, or on more experienced and less experienced ones--in terms of such selection properties. Thus, experimental manipulation of characteristics of instruction--e.g., through creation of two versions of a film or other instructional instrument--can be studied in relation to selection or population variables. This may provide information about the generality of the results for two or more sub-groups of the population, or may provide the basis for identifying differential effects among population sub-groups. The latter may be discovered after the fact, or may have been predicted by a contingent hypothesis that made a differential prediction about the effect of a treatment on two sub-groups of the experimental population.

Film variables (or, more generally, media variables) generally represent the primary independent variables that are of intrinsic interest as experimental variables in media experiments. They may relate to content, manner of presentation, or both.

"External variables" (or procedural or utilization variables) may be manipulated, either singly or in interaction with media variables, so as to reveal the effects of these variables as such, or to investigate their influence on the effects of media variables. Such variables could include conditions of use, special instructions given in conjunction with a program, environmental conditions such as room temperature, and the like. In addition, time can be considered an external variable when, for example, it represents the interval between instruction and testing, and thus permits assessment of the effects of a particular media variable (or another procedural variable) not only in terms of retention immediately after completion of the instruction, but also after a longer period.

As will be noted in relation to procedures for fostering student response to instructional programs (see Chapter 13), and also in the case of other instructional variables, some instructional characteristics may be implemented *either* as media variables or as external variables. For example, question-and-answer pauses might be provided by the instructor between segments of an instructional film, or they might be built into a film. Similarly, attention-directing indicators or pointers might be built into a film by means of overlay animation, or the teacher might perform the pointing and indicating functions during the showing of the film. Thus, from the standpoint of instructional psychology, some variables may be implemented *either as media variables or as utilization variables* (external variables) with no difference in their psychological status. Other variables are, however, by their nature, necessarily either media variables (e.g., the use of color versus black and white in an instructional film) or external variables (e.g., time of day or classroom temperature).

As a way of relating these distinctions to the meaning of an experiment, we may say that the purpose of an instructional-media experiment is, at a minimum, to investigate, and perhaps to establish, a causal relationship between one or more experimental variables and one or more dependent variables. The experimental variables in an experiment on instructional media will always include at least one media variable or one external variable. Thus, one may study the effects of a particular media characteristic, or the effects of a given kind of program, as related to conditions of use (an external variable). In some cases the media variable may be simply the presence or absence of the program. The independent variables intrinsic to the hypothesis may (but do not necessarily) include also one or more selection (sorting) variables. For example, one's hypothesis might be that a given media variable would make more difference on one population sub-group (say boys, or "underachievers") than on some contrasting sub-group.

*Wider and narrower usages of "independent variables"*. The term "independent variable" is often used in a more general sense which includes all relevant as well as all extraneous sources of variation. This wider use would include *all identified or unidentified* sources of influence, including those which are *extrinsic* to (not involved directly in) the hypothesis in which the experimenter is interested, as well as those specified by, and thus of intrinsic interest to the hypothesis under investigation. For example, in testing the effects of a film the investigator might want to consider extraneous factors such as the temperature of the room or fatigue, which might affect students' performance on a "before" test or on "after" test. In a good experiment, *E* (the experimenter) attempts to "hold constant" or otherwise *control* all extrinsic (extraneous) independent variables, so as to isolate the effect of the experimental variables (or variables) *intrinsic* to the purpose of the experiment--that is, the variable (or variables) specified by the hypothesis which is to be tested in the experiment.

The term "independent variable" may also be used more narrowly, so that it is restricted to include only the variables of intrinsic interest to the experimenter, e.g., those whose hypothesized influence he wishes to observe. Generally the sense in which the term is used is evident from the context. When this seems not to be the case, the modifiers extrinsic and intrinsic can be used or, in the case of manipulated variables, we may differentiate experimental variables from control variables (see later classification of independent variables in Chapter 7).

### Variability and chance

All human performances, and measures of human competence or behavior, are characterized by a certain amount of unaccounted-for variability. Two children, even of apparently almost identical backgrounds, behave differently in a given situation; the performance of a given individual on a test will vary from day to day without identifiable reasons. Such unexplained variations are among those that may be attributed to "chance." Their existence poses a fundamental requirement in all educational or psychological experiments--indeed in all educational or behavioral measures, whether in an experiment or not. The requirement is the need to assess the extent to which any obtained degree of difference between two treatments may be attributed to chance variation rather than to actual difference in the effect of the treatments.

Of course this classification of variability is not limited to educational experiments or others dealing with human behavior. A comment in this respect made by Fisher (*Design*, Sec. 31, p. 68, 4th edition) deserves quotation here:

> It is an essential condition of experimentation that the experimental material is known to be variable, but it is not known, in respect of any individual, in what direction his response to a given treatment will vary from the average. No direct allowance for this variability can, therefore, be made. The knowledge which guides us in increasing the precision of an experiment is not a knowledge of the individual peculiarities of particular experimental units ... [e.g., students or classroom groups] but a knowledge that there is less variation

within certain aggregates of these than there is among different individuals belonging to different aggregates.

Memory as related to the requirements of experimental design

To a much more crucial degree than is true in manipulation of physical objects, past experience affects present state. We can heat a copper rod, cool it off to the starting temperature and still be dealing with the same rod. But we cannot "cool off" the effects of experience on behavior in the same way; memory is not really erasable, even if fatigue may be. Once a child has seen a film, $X_1$, on a given topic, we cannot again use his learning to assess the effect of a similar film, $X_2$, on the same topic. This means that in educational experiments, a comparison of the effects of two methods of teaching must either: (a) compare their effects, on different individuals, if they are exemplified in the same subject matter; or (b) compare their effects as exemplified in different subject matters (or learning tasks) if the same individuals are to be treated by both methods (treatments).

The latter (b) is sometimes a very efficient procedure, as it reduces the contaminating effects of individual differences in ability; but it also involves special problems--e.g., that of avoiding having the results affected by the order in which the treatments are presented, and that of assuring comparability of the sample of subject matters to which the treatments are applied. At present we shall confine our attention mainly to the "a" pattern of experimentation, though many of the questions we must discuss will apply also, with some translation, to the "b" pattern of experimentation.

Since (following pattern "a") we are comparing the effects of treatments $X_1$ and $X_2$ on different individuals, the way in which the individuals assigned to $X_1$ and $X_2$ may differ, *other than* having received $X_1$ or $X_2$, is a basic source of possible error in attributing differences in the observed dependent measures to the causal factors represented by the treatment differences. Many of the problems with which we must deal in experimentation on educational media stem directly from this fact. Others stem from inadvertent differences in treatment that somehow tend to creep into experimental procedures, such that we are really comparing $X_1 + Z_1$ versus $X_2 + Z_2$ when we think we are comparing only $X_1$ versus $X_2$. Many of the problems in the experimental "control" of experimental treatments arise in how to deal with, or rule out, such possibilities.

## Some further comments on correlation and causation

"Correlation" means the tendency for two variables to be associated, such that the value or status of one of them (defined, *pro tem* at least, as the dependent variable) can be predicted from knowledge of the other (independent variable), with better than chance success. Correlation of two variables is commonly measured by an index or coefficient of correlation which (see any elementary statistics text) ranges from *zero* (no predictability; variables uncorrelated) to *one* (perfect or one-to-one correlation, in that one variable is completely predictable from knowing the other). In this book we are generally not concerned with the computation or use of correlation coefficients as such, and the term correlation is usually employed merely in the more general sense as just characterized.

Correlation may involve a quantifiable relationship, or may be limited to an unquantified statement of "going togetherness." For example, we may speak of exposure to a film being correlated with amount known about the subject matter, meaning that those who have seen the film know more about the subject matter than those who have not seen it.

It is of crucial importance to recognize that *correlation of variables never, by itself, establishes causation*. For example, as we have previously emphasized, the fact that having seen a film is correlated with knowing more does *not* show that seeing the film was the *cause* of the greater knowledge. It may be demonstrably true that those who have seen it do know more; and this means that by knowing whether a person has seen it we can predict his knowledge with some degree of accuracy. But the only way in which we can really establish that seeing the film was the *cause* of the knowledge is to arbitrarily administer the film in one instance, and withhold it in another instance.

*Confounded variables*. In any comparison other than such an experiment, we do not know what *other* independent variables are *"confounded"* with the particular independent variable that we are calling attention to. In our example, confounded with the independent variable of having seen the film, might be selection variables such as prior knowledge, higher intelligence, or anything that might be the cause or partial cause of the differential status of the dependent variable, namely, knowledge about the subject.

As has been previously indicated, it is a *fundamental characteristic* of an experiment that the experimenter *cause* the variation of at least one independent variable by arbitrarily introducing or manipulating it. This is a fundamental difference between being able to make a rigorous demonstration of the effects of something and not being able to. To say that *only* in an experiment can we say that the independent variable is the cause of the outcome is perhaps going a bit too far. But for most practical purposes in the field of educational media, we can assess the effects of an educational instrument only in an experiment because only in an experiment is the independent variable isolated so as not to be seriously confounded with other possible causal factors. By contrast, correlational or "ex post facto" methods of relating response to exposure fail to distinguish actual effects from *selection factors*, correlated with exposure, which also influence response.

We will use the term *correlational study* to designate a study in which variations in a dependent variable are not demonstrated to be caused by associated differences in an independent variable, because the condition of arbitrary, unbiased assignment of treatments, which is needed in order to demonstrate such a cause-effect relationship, is not present. In a correlational study the independent-variable categories are *taken as found* rather than being arbitrarily administered; such a study can demonstrate that an independent variable is correlated with, or can predict, a dependent variable--but does not demonstrate that the first is the cause of the second. It should also be emphasized that correlational results obtained in an experimental context do not thereby acquire the causal validity of experimental results, which attaches only to conclusions about those factors that were in fact manipulated experimentally.

### *Experimental-manipulation versus correlational comparison*

We have stressed that the effects of an instructional instrument's use can usually be established only through measurements made in connection with the controlled administration of that instrument. One basic aspect of being "controlled" is that the instructional instrument is administered arbitrarily to a given group of students, in order to reveal the effects it produces--without the contamination of spurious or confounded factors which could lead to erroneous attribution of "effects" to the instrument. The primary basis for ruling out merely correlational data as a sound basis for identifying the effects of instructional media is, as has been implied, that doing so inevitably leaves open the confounding effects of biased selection (including "self-selection"), and related sources of contamination, thus confusing correlation with causation.

The validity possible with the experimental method is thus contrasted with survey techniques which attempt to get around the inherent fallacy of correlational methods for determining "cause and effect". It is also contrasted with unvalidated questioning techniques which essentially consist of "invited introspection" and depend on accepting a respondent's statement as to what caused his behavior. Many examples could be cited where the results of using such techniques were confidently accepted until later application of experimental method showed that expectations based on them were quite erroneous.

For example, interviews or questionnaires given to those emerging from a theater where a film on anti-semitism had just been shown might show that the degree of anti-semitism was much less than among the general public as a whole. However this would by no means demonstrate that the film had the effect of reducing anti-semitism, because those who chose to see the film might well have done so because they were already disposed against anti-semitic views. Similarly, if we find that readers of *Time* magazine knew more about the international situation than non-readers, this does not necessarily mean that all (or indeed any) of their superior information was gained and retained from reading the magazine. Those who decide to read the magazine are very likely better informed to begin with, and would have gotten higher scores even if the test of information had been given before they did the reading instead of afterwards.*

At the risk of possibly unnecessary redundancy, it is therefore once more reiterated that the fact that those who have studied a particular program know more about the content than those who have not obviously does not mean that the difference in knowledge is the result of having used the program. This can be safely concluded only when the program was arbitrarily administered to one group of individuals and withheld arbitrarily from another group. Otherwise, there is no way to preclude the likely possibility that those who used the program already knew more about the content than those who did not, and/or that the observed superiority results from the fact that they chose, or were chosen, to use it because of some selection characteristic such as greater interest in, and correlated knowledge about, the subject matter. If we seem to reiterate this point ad nauseum, it is only because the point is so frequently overlooked in attempts to study the effects of media, even by otherwise quite sophisticated investigators.

Despite many ingenious attempts at matching and equating to insure that those who had the program are "otherwise comparable" with those who did not, there is in general no way to rule out the adventitious correlates of self-selection, even though the groups compared are made "comparable" with respect to a host of selection variables. Thus, we reiterate that the first essential for the experimental identification of the effects of an instructional instrument is that of arbitrary administration of the instrument. The word arbitrary, as used here, implies, for rigorous experimentation, some form of randomization. The crucial role of randomization in eliminating systematic bias from experimental "effects," and in providing a defensible, unbiased estimate of error, will be further discussed presently.

### Control

The term *control*, as applied to an experiment, means any procedure which is introduced to make the interpretation of an inferred relationship between a cause (an independent variable or experimental treatment) and an effect (instructional outcome, result of dependent variables) as unambiguous as possible.[*] More specifically, "control" refers to any techniques which foster the attribution of effects to the defined experimental variables rather than to alternate possible variables.

Factors which may contaminate, or bias, or make ambiguous the interpretation of a causal relationship between independent and dependent variables are generally controlled by trying to insure constancy of conditions, by some form of counter-balancing, or by randomization. Generically, control includes randomization; however, in some instances, control, in a more specific sense, is used to designate controls other than randomization (See Chapter 5).

Control does *not* necessarily imply "a control group". The concept of control of extraneous sources of influence by control procedures is reflected in the popular conception of "holding constant" all factors except the designated experimental variable. This conception, though it is an oversimplification and hence can be misleading, is not a bad first approximation as a way of conceiving the function of control in an experiment.[*]

## Control, bias and randomization

The question of "control" is so important in experiments that some individuals use the words "controlled experiment" as one word (just as some Southerners allegedly use the words "dammyankees" as one word). However, although the aspect of control is one about which we will have a great deal to say, it should be recognized at the outset that one can never, in any direct sense, control all, or anything like all, of the elements that go into an experiment.

The fact that one cannot do so is a reason why randomization is a concomitant of a successful experiment. Because in practice we can't actually control everything in the experiment except the one thing that we are trying to study, we try to arrange conditions so that all of the adventitious things that might affect the result are "randomized out," so they do not systematically bias the results of the experiment. The one thing that is systematically introduced is then the specific factor that we are trying to study. We shall return frequently to the notion of randomization; it will be illustrated in an experiment described in Chapter 4 . For the present it will suffice if we understand that factors which cannot otherwise be fully controlled can, by insuring that their influence is left strictly up to the operation of chance, be prevented from systematically biasing the results of an experiment in either direction.

## Experiments as tests of hypotheses

One kind of statement about experiments says, in effect, "An experiment comprises a situation devised to test a hypothesis." Though this is true and very important, we will also want to think about some uses of experiments where the hypothesis to be tested is not as clearly evident as in other instances. Still, for a working definition, "a situation devised to test a hypothesis" comes pretty close to the crux of what we have in mind, when variables specified by the hypothesis are manipulated in a controlled fashion by the experimenter in order to determine the tenability of that hypothesis. In Chapter 5 we will examine further the nature of scientific hypotheses, and the conditions for valid tests of their tenability. For the present we may define a hypothesis as any proposition which might be advanced to explain an observable event or a class of observable events--in our case, the effects associated with the use of instructional media. An experimental hypothesis is a hypothesis which an experimenter is interested in verifying, stated in such a form that it may be subjected to experimental test, or that a logical consequent of the hypothesis of interest can be stated in such a form that it may be subjected to experimental test. Alternate hypotheses (see Chapters 5 and 6) are other reasonable explanations for the occurrence of the observations predicted in the experimental hypothesis from which the experimental hypothesis must be differentiated.

## The null hypothesis...

The so-called "null hypothesis" is a particular alternate hypothesis, always needing to be considered explicitly in any educational experiment, which states that the alleged "effects" observed following an instructional program can, in fact, be explained solely by the operation of chance factors and thus are not attributable to the instructional program. In the logic of statistical inference underlying most experiments on instructional media and educational methods, as well as experimentation in the behavioral and biological sciences more generally, data are ordinarily analyzed with reference to a "null hypothesis" which asserts that two or more samples of data are in fact random samples from a common population. That is, the null hypothesis supposes that the independent variables introduced in the experiment had no systematic effect, so that there are no differences between the data that result from two or more experimental treatments (e.g., two forms of instructional programs), except those readily attributable to chance.

The experimenter then proceeds to test the tenability of this null hypothesis. He does so by reference to appropriate tabulations or calculations which indicate the probability that, _if_ the null hypothesis were in fact true, differences as large and/or as consistent as those he observed could occur simply as a consequence of chance or random sampling fluctuations. If this probability is lower than some value (such as .01) agreed on conventionally as an acceptable level of statistical "significance," he rejects the null hypothesis. In doing so, he concludes that it is preferable to accept the alternative conclusion that the observed difference resulted from some non-random causal factor. This factor, in a well-controlled experiment, he will normally indentify as the difference in the experimental treatments he employed. The logic underlying this argument is illustrated in Chapter 4 and also will be further commented on in Chapter 5 and later chapters.

Illustrative experimental hypotheses. Here are some examples of hypotheses concerning educational media which we might want to test by experiment.

(1) Seeing a particular film will teach how to tie a bowline.

(2) A color film on a particular subject will teach more than a black-and-white film.

(3) A demonstration film that shows a procedural task to be learned will teach better if the task is filmed so the camera "sees" the performance of the task from the viewpoint of the performer rather than from the viewpoint of an observer watching him perform the task.

(4) A self-instructional program in which the student has to write his responses to each item will produce better learning than if the student merely answers the questions mentally.

(5) A televised lesson on high school chemistry will stimulate more interest in further study of chemistry if televised experiments are performed by a student rather than by an adult instructor.

There are some serious difficulties and ambiguities in each of these illustrative hypotheses, the nature of which will be more apparent later if they are not apparent immediately. Overlooking such difficulties for the present, let us take any one of them at face value and see how we might try to test it by an experiment.

Testing an illustrative experimental hypothesis

As our example we will take the second of the above hypotheses--i.e., the hypothesis that a color film will teach more than a black-and-white film. We will not at present concern ourselves with whether this is a sensible hypothesis nor whether the results could be generalized. Nor will we ask, if the color film proves to be better, what the reasons for its superiority may be. Finally, for the present, we shall accept as the sole relevant dependent variable the amount of factual information learned from the film, as revealed by a relevant test of information; and we shall assume that this test is already available.

In trying to study the effect of color versus black-and-white motion pictures, our defined instructional objective might be to increase favorable attitudes toward the ideas of international cooperation, or to teach ability to recognize and differentiate the fauna and flora of Central America. Let us say the latter is our objective.

We are at once faced with some interesting problems in trying to conduct such an experiment. Basically, of course, the paradigm for such an experiment (the pattern of basic design) might be something like this:

Show color film to group I;

Concurrently show black-and-white film to group II;

Test both groups to see what they have learned.

Suppose we do show one group of people a black-and-white film, simultaneously showing another group a color film, and then give both groups a relevant "test." Then we look at the test results and try to see whether we can draw a valid conclusion about the relative merits of color and black-and-white films.

The essential notion behind this experiment, following the paradigm (which will have many elaborations as we go on) is a simple enough one. What we want to do is find out what people learned from the black-and-white film as against the color film. To this end we try to control as many sources of influence as possible, _except_ for the one factor we are manipulating. What are some of the obvious things that would need to be controlled? Subject matter, perhaps? We wouldn't get very far if we used a black-and-white film on the _fauna_ of Central America, and a color one on the _flora_ of Central America, and tried to compare them with some kind of a test and see which one was better. (We are not saying this would be impossible, but it would be extremely difficult and it would be asking for all kinds of unnecessary trouble.) So let us assume that the first thing to control is at least the subject matter of the film; probably we would also control the details of content. A straightforward way to do this (although some arguments might be raised against doing it in just this way) would be to take a picture that was shot in color and make as good as possible a black-and-white print of it. We might then compare those two, with the assurance that everything in the content and presentation of the two films was identical except for the differences associated with the fact that one was in some kind of color, and one was in black-and-white. So at least we would have ruled out most of the grossest sorts of adventitious differences in presentation.

Now we must try to show the two films to two "comparable" groups of people. (We can't use the same subject matter twice on the same group of people.) So let us say we will use two different groups of children, and thus we have to know what _kinds_ of _different_ groups of children we are to use. Assuming that our subject matter is appropriate for the children, it is customary to say that the two groups have to be "comparable" and that they have to be "representative." These two words are perhaps more misunderstood and more subject to unfortunate misinterpretations than any other terms in educational research.

### Representativeness

When we talk about experimental conditions (including the learners) being "representative," we mean in the first place, at least that the learners to whom we show these films, and whom we then test, are not grossly atypical of the kinds of people with whom we hope to apply the results. If, for example, we are interested in using color films for seventh-grade children, we presumably would not test the film on college students, and we certainly would not test it on people who were color blind. Thus, in some gross sense at least, it is obvious what we mean by saying that we want to have the sample group of children chosen so that it is representative of some larger group it is supposed to represent. How completely we try to define the population that our sample is supposed to represent, and the degree of care exercised to insure that our sample is in fact representative, can be of varying degrees of importance, depending upon how careful and how cautious we wish to be. How much attention we pay to such matters may depend also on the kind of conclusion we are trying to draw as we shall see later on.

### Comparability

Second, and more crucial, we must insure that the groups are "comparable." One way of saying this is that they *both* have to be representative of the *same* population. In other words, the two groups, which will never be quite identical, must be "substantially equivalent." More exactly, any differences between the two groups have to be such that two conditions are satisfied.

First, any differences between the people who are viewing the films--the color and the black-and-white one--*must not systematically bias the outcome* of the experiment in favor of one treatment or the other. (We must say that the differences must not *systematically* bias the experiment, because we can never say that they might not influence it to some extent.)

Second, the make-up of the two groups of people has to be of such a nature that the amount of *unsystematic* influence (resulting from the fact that inevitably there will be *some* differences between the two groups) can allow us to form *an unbiased estimate of how much the results could have been affected by these differences between the groups*. In other words, we must be able to calculate a *valid measure of the error in our experiment*. Indeed, the ability to calculate a valid measure of error is the reason that we introduce randomization, and is a *sine qua non* for a valid experiment. Remember that we never can control everything, so in an experiment we always must have the ability to *assess the amount of error* that is present. It is only by comparing our results with such an assessment of the errors to which the experiment is liable that we can ever state whether we have arrived at a defensible conclusion.

Comparability can never rigorously be established after the fact by comparing the characteristics of the two groups. It can only be rigorously assured by a procedure, involving random assignment and control of experimental treatments, which allows us to calculate a valid estimate of experimental error.

### Randomization

Randomization means the assignment of subjects, tasks, conditions, procedures or any other element in accordance with a random procedure. A random procedure for selection or assignment of things (e.g., the elements that enter into an experiment) is a procedure such that each element has an equal chance of being selected on any given occasion; that is, the selection is determined purely by chance. (Strictly speaking, random can be considered to mean that all combinations of items or elements are equally likely; but for our purposes it will suffice to consider random to mean that every item or element is equally likely.) In a strict or rigorous randomization procedure, this is insured by use of some explicit procedure known in advance to accord closely with the theoretical laws of chance, preferably by the use of authenticated tables of random numbers. However, substantially equivalent results can generally be obtained by any "chance" operation such as coin-tossing, dice-rolling. Indeed, for some purposes assignment in accordance with any unsystematic number series (e.g., last two digits of a soldier's army serial number, or the last two digits of successive numbers in a telephone directory, in which an extended series generally is not purely random) would closely enough approximate conditions of randomness for many purposes, since it should at least remove any systematic bias. But tables of random numbers are preferable and, since they are readily available, their use is to be recommended whenever possible (see Appendixes A & B). As we will see, it is also generally necessary to consider some form of *random distribution*. This is a distribution of frequencies of occurrence that is generated by (or corresponds to what would be obtained by) purely chance occurrences. The normal or Gaussian distribution (which occupies such a central role in most statistics textbooks) and the binomial distributions are important special cases of random distributions. Their importance for our purposes lies in the fact that they are frequently used as the basis for *statistical tests* to determine the likelihood that a given experimental result could arise purely as a chance occurrence.

### Purpose of randomization

The purpose of randomization in experimentation is that, by proper randomization, we are potentially enabled to insure that *all* of those factors which we could not control experimentally or hold constant through procedural means, and which, therefore, will have some undetermined effect, have this undetermined effect in such a way that two things are true. First, that the effect *cannot systematically bias or prejudice the outcome* of the experiment one way or the other (there are no prejudices in a table of random numbers). By randomizing, we therefore see to it that both treatments get "a fair shake."

The other crucial and related consequence of randomization is that it enables us not only to insure lack of bias due to adventitious factors, but also to measure or form a *valid estimate of the amount* *of error that we have in the experiment*. This permits a so-called statistical test of "significance"--a basis for assessing how dependable the results are, or how much confidence we can place in them. Basically, this test is made from the standpoint of being able to say how often we would get a difference as large or larger than the one we did get, just as a result of chance.

It is *only* by introducing randomization at some appropriate point in an experiment that we can ever get a valid test of "significance" or confidence, because the so-called significance test is in fact nothing more nor less than a way to get a statement of how often would there be, by chance, a difference as big--or, more exactly, as consistent--as the one we got. Thus the essential question that is asked is: Was the difference we got sufficient enough that it would occur *so infrequently* by chance that we are willing to throw out the supposition that it did represent just chance fluctuation? The only way we can make such a statement validly is to explicitly *introduce* some element of randomization, that is, to let chance operate explicitly in the experiment at some point. This is a truth that was not generally recognized until it was pointed out by R. A. Fisher;* it will be further elaborated in the next chapter.

There are many different ways to randomize, and the *way* that we would do this may be affected by the varying circumstances of different experiments; but the *necessity* for randomizing is invariable if we are to have an experiment whose validity can be rigorously defended. That is, we *always* have to randomize at some point (and sometimes in more than one way) in order to have an experiment that is an experiment in the true or formal sense of the word, i.e., one from which we can draw conclusions about cause and effect with an explicit, identifiable rationale, with the concomitant ability to defend the validity of the result (rather than merely having a basis for being "convincing" in an informal way). It should also be made explicit that randomization itself is no guarantee of the validity of experimental results; this guarantee occurs only if there is a logical one-to-one relationship between the acts of randomization and the manner in which the estimate of experimental error is made. The necessity and importance of this statement will be made clearer through an example in Chapter 4 , and its implications will come up several times again later--e.g., relative to the consequences of the kind of group sampling frequently employed in educational-media experiments. For the present suffice it to say that randomization is no mere ritual to be introduced at some point and then forgotten; many patterns of randomization are possible, but a particular pattern, once adopted in an experiment, must govern the interpretation of the results in relation to an estimate of experimental error, from there on.

Statistical tests

**Statistical tests are an inevitable and necessary part of any formal experiment to test a hypothesis. The nature of specific statistical** tests will be discussed later. Here we merely wish to note that a statistical test is a way of arriving at a probability figure which indicates the dependability of the finding or, to use an unfortunate but commonly used term, its "statistical significance," (see pp.         ). The essential role of a statistical test of confidence is to provide an assessment of whether the observed effects are sufficiently consistent in direction (and, often, in amount), in relation to the variability in our experimental measures, so as not to be reasonably attributable to chance variation--i.e., to be able to reject the null hypothesis. As noted above, this requires a defensible method of assessing experimental error, in order to demonstrate the non-chance character of the differences that purport to reflect the effects of an instructional instrument (or the differences between the effects achieved by two instruments). This demonstration of "statistical significance" or, to use better terms, "confidence," or "reliability," or "dependability" or rejection of the null hypothesis, is a necessary but not a sufficient condition to support the claim that the effects of an instructional treatment have been revealed. Thus, effects of a program are indicated only if the "significant" (or dependable) difference between measurements is obtained under the conditions of a purely controlled experiment, as characterized above.* This fact will become clearer when we further consider the status of the null hypothesis as merely one of a number of alternate hypotheses which might explain any given observed outcome in an experiment (cf. Chapters 5 and 6).

Certain requirements for valid and efficient tests of "significance" (i.e., dependability) will be further discussed in Chapters 4, 8, and 10.

## Some Characteristics of Experiments

### What makes a good experiment?

There are only really two central questions that determine the basic worth of an experiment. First, did it attempt to answer an important question? Second, did it produce a valid answer to the question? The first question is by far the hardest to answer affirmatively; how to get a good idea or pose an important question generally defies analysis. Such suggestions as can be given will come largely from three sources: (1) after-the-fact realization of the futility of certain kinds of experiments even when carefully executed; (2) a review of the research literature to identify what has been found out and where the major gaps of ignorance seem to lie; and (3) a consideration of two kinds of theory--those dealing with the nature of learning or instruction and those dealing with the rationale of scientific inquiry. The latter of these kinds of theory largely determines, also, the answer to the question of validity, as will be further elaborated in the next three chapters--first by examining an illustrative experiment (Chapter 4 ), and then by a further consideration of the logic of hypothesis testing (Chapters 5 and 6 ).

Critical characteristics

Experiments can be and are criticized on many bases, some reasonable and some unreasonable. The purpose of this section is to identify a set of characteristics that are useful for discussing, evaluating and criticizing experiments. These are relevant to the experimenter himself in controlling and directing his own research efforts, and for pointing out to his readers in his research reports those aspects of his work which are specially relevant to their consideration. These are also proposed as the frame of reference for evaluating an experiment or a proposal to do an experiment. In this chapter these characteristics will be introduced and defined. They will be given more extensive discussion in several later chapters (particularly Chapters 10 and 11).

Salient characteristics or aspects of an experiment are: the importance of the experiment; the validity and generalizability of the experiment; the sensitiveness of the experiment; the precision of the experiment; and the efficiency of the experiment.

Importance. Importance refers to the characteristic that has most primacy in considering the worth of an educational-media experiment, namely, is the study addressed to a significant question? Is knowing the answer to the question, or testing the hypothesis the experiment proposes to test, or establishing the conclusions it could potentially establish, likely to make a significant impact on the science of instructional media? As we have indicated, the question of how to determine what is an important experiment is a most difficult one. What little can be said about this will be discussed primarily in

Chapter 12. Here we wish mainly to point out that an ingenious, methodologically impeccable experiment can be addressed to a trivial question. However, we want to point out also the need to distinguish between the importance of the question itself and the validity of an experiment (or its rigor), as determined by the extent to which it complies with formal requirements of valid experimentation.

The importance of an experiment will depend both on the characteristics of the independent variable that is to be introduced experimentally, and also on the nature of the relationship hypothesized between this experimental variable and the dependent variables it is presumed to affect. Comments on the latter--the nature of the relationship hypothesized--will be deferred to later chapters (Part II). With respect to the independent variable, we can say that its potential importance will depend on its inherent potency, on the degree of generality with which it can be defined, and on the pervasiveness of its applicability in instruction.

By potency we mean how strong an influence the factor is capable of exerting. In a sense it is a purpose of experimentation to determine this; hence some circularity is involved. To some extent, that is, the importance of an experimental variable is to be assessed by the results of the experiment rather than pre-judged in advance. On the other hand, both a priori considerations and empirical evidence from past related experimentation can help in estimating the probable potency of an independent variable in instruction. For example, there are at present rather strong reasons, on both empirical and theoretical grounds, for thinking that repetition is a potent factor and that prompting techniques (for insuring the correctness of the responses a student makes during practice) are also potent factors in determining the effectiveness of instruction. Hence, experiments designed to test the relative effectiveness of different ways of implementing these factors seem likely, *a priori*, to be dealing with potent variables that are potentially capable of making a difference of considerable consequence in the effectiveness of instruction. On the other hand, there is at present little reason to suppose that, within wide limits, the exact style of type and shade of ink in which verbal material in an instructional program is printed are likely to be very potent factors in determining the outcome of instruction.

However, even though a factor may be a potent one potentially, it may not be implemented in an effective or efficacious manner in the experimental treatment. *Efficacy* of implementation is, therefore, to be distinguished from potential or inherent potency of an independent variable (see p. 7, Chapter 7 ). In line with this distinction, *potency* helps to determine the potential importance of the experiment, whereas efficacy affects its *sensitiveness*.

The question of *generality of definition* is related to the *reproducibility* of the independent variable; this characteristic is obviously one of the factors that will affect the extent to which the results of the experiment can be generalized to other situations outside the immediate context of the experiment itself. This characteristic of generalizability also has a relation to the validity of the experiment as will be noted below. The question of generalizability is a most difficult one; some of the considerations that affect it in terms of the definition of independent variables will be further discussed in Chapters 11 and 12.

To illustrate the quality of pervasiveness of applicability of independent variables, it might be considered that the use of slow-motion photography could be a potent factor, but perhaps only in the teaching of a certain rather restricted class of skills. From this standpoint, hypotheses concerning its influence might in a sense be considered somewhat less important than those dealing with characteristics of, say, repetition, which might be expected to be more widely applicable in many different forms of instruction.

Validity. Generically, the concept of validity indicates whether something truly represents what it is supposed to represent, without bias or contamination from extraneous sources. Validity does not necessarily imply precision, but only absence of bias.

A differentiation in meaning is necessary between validity as applied to a "test"--in the sense of a measuring instrument, such as a test of aptitude, achievement, or intelligence--and as applied to an experiment (the "test" of a hypothesis). A test--i.e., measuring instrument--is said to be valid if it measures the attribute or characteristic which it is alleged to measure. An experiment (to "test" a hypothesis) is valid if it permits valid conclusions; i.e., an unambiguous interpretation of the evidence with respect to the hypothesis which the experiment was set up to test, or a valid measure of the effect which it was designed to reveal.

A distinction is also made by some (e.g., Campbell and Stanley, 1961) between "internal validity" and "external validity." "External validity" is roughly synonymous with generalizability of results--will they hold up in other situations to which we wish to apply them? The question of "internal validity" is a narrower one--it asks only whether, within an experiment, the evidence provides a valid basis for the effects observed. Internal validity is thus a necessary condition, but not a sufficient one, to assure external validity or generalizability of an experimental result. Generally, when used herein, without a modifier, "validity" refers to the internal validity of an experiment. The relation of validity to generalizability will be returned to later (cf. esp. Ch. 11). For the present we may note that the internal validity of an experiment is related to the notion of so-called "comparability" (see pp. , supra), while the concept of external validity or generalizability is related to the notion of "representativeness" (cf. pp. ).

The notion of validity in an experiment implies, as stated above, absence of bias. By bias we mean any extraneous process, occurrence, or procedure which prejudices the outcome of an investigation (unintentionally or otherwise) in such a way as to favor one group or treatment or condition in a systematic fashion not intrinsic to the experimental variable as defined. Insurance against bias is obtained either through procedural control to achieve indentity or comparability of conditions, or through randomization. Bias is what may impugn the internal validity of an experiment: (For example, in comparing the effectiveness of two films, one might consistently be used with better projection conditions, or shown to more able students, than the other.)

The validity of an experiment is, then, insured by seeing to it that potentially biasing factors do not operate so as to systematically or consistently favor one treatment over another. This is always done partly by excluding or "holding constant" such biasing factors (through procedural control or by balancing out their effects), and partly by randomization procedures which, while not removing or cancelling the effect of the biasing factor, allow it to operate only within limits set by the laws of chance.

Sensitiveness. Sensitiveness can be defined to mean the likelihood that, all things considered, the effects postulated by an experimental hypothesis will be revealed in a valid experiment if the hypothesized relationship predicting them is in fact true. Sensitiveness is affected not only by precision (see below) but also by the efficacy of the experimental treatments.

Often the term sensitiveness appears to be used interchangeably with precision (e.g., Fisher, Secs. 11&12, pp 21-25). However, it is convenient to make a distinction between the two, particularly in the case of educational experiments. The distinction is that, in referring to precision, we talk about only those things that affect the size of the estimated experimental error, whereas by sensitiveness we mean the likelihood that a particular effect will be revealed. Precision contributes to sensitiveness, but so do other things. A major factor contributing to sensitiveness is how likely we are to get a large difference. This in turn means the extent to which a "green thumb" has been applied in implementing the experimental variable--how efficaciously we have implemented the experimental factor under consideration. It seems desirable to think about sensitiveness as a separate kind of characteristic concerning which to inquire about an experiment, even though on the one hand, sensitiveness inevitably gets into the question of validity and, on the other hand, gets into the question of precision (since sensitiveness is affected by precision). The factor of "efficacy" of experimental treatment is especially a cause for concern in educational experiments (e.g., we may introduce a "motivational" factor that actually turns out not to be an effective motivator). It is less clearly applicable, or at least is less likely to be a source of worry, in physical, biological or agricultural experiments, (where the independent variables can be more exactly, and thus more reproducibly, specified). But in educational experiments we always have to inqure whether the experimenter may not have implemented the experimental variable with sufficient skill or extensively enough to make the resulting treatment an efficacious factor in his experimental program.

Precision. By this we mean essentially the power of an experimental comparison, that is, its capability for detecting a minimum size of difference in the dependent variable as a non-chance effect. A distinction should be made here, however, between the precision of the measuring instrument and the precision of the experiment as a whole. Greater precision of the measuring instrument may be achieved in various ways (see Chapters 8 & 11)--for example, by increasing the number of items or questions in a test. The result is to yield an instrument with a smaller standard error of measurement. A more precise measuring instrument will, other things being equal, contribute to the precision of the experiment. However, so will other things, including the degree of control exercised, the structure of the experiment, and the size of samples that are used.

It should be re-emphasized that the precision and the sensitiveness of an experiment affect the economy or efficiency of experimentation; as such—they do not affect its internal validity. An insensitive experiment can provide a valid test of a hypothesized effect if the effect is a potent one. Lack of precision merely means that the effect may not be revealed even though it is in fact present; a larger or otherwise more precise experiment might reveal it. Similarly, insensitiveness due to inept implementation of a hypothesized factor may mean that its effects are not manifested, even though with more efficacious implementation the effects hypothesized might show up even without increasing the experiment's precision.

In general, we can say that the result of increasing the precision of an experiment is to reduce the amount of variation, in the dependent variable, that is attributable to chance error—and hence, to increase the likelihood that an effect of a given experimental treatment will be revealed as a demonstrable, non-chance effect. But the experiment can be perfectly valid—even with a very large amount of chance error—provided only that this error is, in fact, properly assessed in the experimenter's estimate of error, whereby we test the dependability ("significance") of our results. The precision of an experiment will obviously be related to the heterogeneity of the population of learners used in the experiment. Ways in which error can be reduced, for a given degree of heterogeneity, include the use of stratifying, matching or pairing, as well as the procedural control of environmental sources of unnecessary variability. Such procedures for improving precision will be further discussed in Chapter 11 . For the present, our main purpose will be accomplished if the reader fully grasps the distinction between the essential conditions of validity in an experiment and the desirable but not essential condition of precision.

Great precision is required only where very small effects (or very small differences in effects) must be detected. When this is not the case, a quite satisfactory experiment may be fairly insensitive or imprecise provided the conditions for validity are met. Finally, it may be noted that in a technological or evaluative experiment in which we are concerned only with gauging the question of whether a particular program treatment had a given effect (and not with testing an hypothesis of more generality about the effects of a kind of treatment), the terms "imprecise" and "insensitive" (or precise and sensitive) can be applied interchangeably. In this case, we are not concerned whether the treatment was an efficacious implementation of a hypothesized factor (since the "hypothesis" is merely that this treatment had an effect) and thus, as defined, the sensitiveness of the experimental test is wholly governed by its precision.

**Efficiency.** The efficiency of an experiment refers to the economical use of the resources of time, money, subjects, etc. It also has a narrower definition in terms of the ratio of the precision exploited to the precision possible in the experiment.

Status of these descriptive categories

To some extent the several above-characterized categories of description overlap. Despite the overlap, it is useful to think of them as separate categories in the development and critique of experiments. The most useful or critical differentiation is between validity and sensitiveness or precision; it is important to see that we can have a valid experiment even with low precision. Concerning some of the characteristics (precision; efficiency), quantitative statments can be made. (For example, some aspects of precision or power are expressible in terms of numerical coefficients such as the "power" or "power-efficiency" of a statistical test.) About others only qualitative statements (or quantitative ones in the sense of greater or lesser) can be made. These two classes of statements might be referred to as "analytic" and "non-analytic" statements.

## Criticism of experiments

As has been implied, not all of the statements that savants make about experiments are really relevant criticisms. More important, from the point of view of this chapter at least, is the fact that criticism leveled at one aspect or characteristic of an experiment sometimes is really relevant to a different aspect. In particular, the most common error is the confusion of the *validity* and the *sensitivity* or *precision* of an experiment. That is, a criticism may be leveled at an experiment which is presumed to put the validity of the experiment in question, but actually is relevant only to the sensitivity of the experiment. Making a statement about the frequency of such criticisms which is impressionistic and not really based on any counts, we would, indeed, say that the most common misplaced criticism is that of attacking the validity of the experiment on the basis of such-and-such a factor's not being controlled, when in fact the factor was controlled, but was not controlled in the way the critic thinks it should have been controlled.

Often it is found that such-and-such was controlled by randomization. We have said that every true experiment *must* employ randomization. Given proper randomization, a large number of factors or variables are automatically controlled whether the experimenter deliberately made a decision to control through randomization or not. A larger degree of dependence for control through randomization, though it is adequate control, does increase the experimental error over what it might have been had some other kind of control been available and used--that is, it may adversely affect the experiment's precision though, at the same time, it may not only not impair but may help to guarantee the experiment's validity.

"Good" and "bad" experiments

A torpid old drunk being wheeled into the emergency room of the General Hospital heard one of the interns say, "He must have gotten hold of some bad whiskey." The old man bestirred himself enough to say, "There is no such thing as bad whiskey. There is only good whiskey and better whiskey."

With little exception, we feel the same way about experiments. The only time we would certainly be willing to say that an experiment is completely useless is when the basic logical requirements for validity are violated or when the question asked is clearly a very trivial one.

Justifiable criticism of experiments on the basis of low efficiency or sensitivity usually does not mean that the experiment was bad per se, but only represents a statement of sad concern that resources were wasted when there is so much ignorance and it is so difficult to penetrate it.

Even when the validity of an experiment is attacked on the basis of bad logic, it is sometimes possible to salvage the experiment by correcting the logic. This may or may not be possible depending on the nature of the logical error and when it occurs in the experiment. For example, the experiment can almost invariably be salvaged if the only logical error is in the last step--the stating of the conclusion.

# Chapter 4
# Illustrative Hypothetical Experiments

The purpose of this chapter is to stress and exemplify certain basic principles that are fundamental in considering the design of any experiment on instructional media--or, for that matter, any experiment at all in which we have to deal with quite variable data. The experiments are hypothetical in the sense that they do not report actual experiments that have been conducted on instructional media; however, they exemplify patterns that could be (and in the case of the second example, frequently have been) used in such experiments. The authors believe that a careful study of the two experiments described in this chapter will well repay the reader, and that the principles exemplified can perhaps be better apprehended as they are exemplified in specific (even though hypothetical) examples than could be done solely through more abstract exposition.

Some of the basic principles and concepts exemplified by the first experiment described in the present chapter are:

(1) The need to analyze in advance, in any experiment, all possible outcomes, and to decide what unambiguous interpretation can be placed on each, should it eventuate;

(2) The essential role of chance in any experiment, and the reasons why explicit provision must be made as to how it will operate in the experimental situation;

(3) The intimate manner in which the structure or pattern of

arrangements of an experiment determines the interpretation that must be placed on its results through a "test of significance";

(4) The desirability that the experimenter (and also the reader of a report on the experiment) gain as clear as possible an understanding of every single step in the logic of the experiment, so that as little as possible need be taken on faith;

(5) The nature of the choice open to the experimenter with respect to the degree of confidence to be achieved before accepting the tenability of an experimental hypothesis;

(6) The basic dependence of experimental validity on an orderly, logical method of procedure, rather than on isolated facts, however accurate these facts may be;

(7) The way in which probability calculations enter into the interpretation of every experiment;

(8) The meaning of the "null hypothesis" in relation to probabilities, and the realization that the null hypothesis can never be proved, but enters into experimental logic only to be possibly disproved;

(9) The crucial role of randomization in experimental operations, and the way in which it, and only it, can guarantee the strict validity of the experiment;

(10) The crucial difference between the essential requirement of validity and the merely desirable characteristic of sensitiveness in an experiment, and some of the conditions that affect each.

## EXPERIMENT 1: Basic Principles of Experimentation as Exemplified in an Illustrative Experiment on Stimulus Discriminability

The experiment to be described here is presented because of the unusually clear insights it provides into some key aspects of the basic logic of experiments. Because of its pedagogic value it seems useful to present it, even though it deals not with the direct effect of media on the intended audience or learner, but rather with the discriminability of stimulus differences in media.

The experiment is based on, and closely follows, a discrimination experiment described by Sir R. A. Fisher in the second chapter of his classic work, The Design of Experiments (see Preface). The original experiment deals with sensory discriminability in tea-tasting, and the statistical test employed is sometimes referred to punningly as the "tea" test (as distinguished from the much better known "t-test" based on the work of William Sealy Gosset--who published under the pseudonym of "Student" and whose test, which represented a major breakthrough in the logic of scientific inference, employs a statistic accordingly known as "Student's t"). In the present description, the basic features of Fisher's exposition have been preserved in an application more relevant to instructional media--namely, the discriminability of differences in the stimulus picture provided by different methods of reproducing an instructional motion picture. Extensive quotations from Fisher's original account have been identified by abbreviated reference citations as explained in the Preface. Extended underlining of more than a word or two in these quotations is, unless otherwise noted, that of the present authors rather than Fisher's.

Statements reflecting a number of general principles, applicable (as stated, or with only trivial modification of form) to any experiment on instructional media, have been thus underlined. (Double underlining italics in the final reproduction indicate words italicized by Fisher).

### Statement of an experiment

A film producer asserts that by looking at a film he can tell whether it is a "reversal print" (obtained by the process used customarily for preparing black-and-white "home movies"), or whether it was made by what he feels is the superior-quality process of printing it from a master negative. We want to consider the problem of designing an experiment by means of which this assertion could be tested. For this purpose we will first describe a simple form of experiment, and will then consider some of its characteristics and its limitations. In doing so we will try to identify those features which appear to be essential to the experimental method, and those which represent desirable or interesting features or methods which, however, are not essential.

For our experiment we shall prepare eight short film sequences, four in one way and four in the other, and shall present these to the subject (our film producer) for judgment, in a random order. We have told the subject in advance what the test will involve—namely, that he will be asked to view eight short films, that these will be four of each kind, and that they will be presented to him in a random order. By this is to be understood not an order determined capriciously or haphazardly by human choice but, rather, one determined carefully by the actual manipulation of physical apparatus used in games of chance—e.g., the toss of a coin or roll of dice. (Actually, since these devices are seldom perfectly unbiased and, in any case, may not be convenient to use, we might better employ a published collection of random numbers, which may be based on the results of such manipulation conducted under as close to ideal conditions as can be achieved.) Our film producer's task will, of course, be to identify the nature of each of the eight films, so as to divide them into two sets of four, which will, if possible, agree with the film treatments received.

In discussing the parallel experiment on which this one is based in terms of "Interpretation and its Reasoned Basis," Fisher presents the following most cogent argument, which applies to this and any other experiment with equal force. Says Fisher (1960, Sec. 6, p. 12):

> In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each of them. Further, we must know by what argument this interpretation is to be sustained.

Fisher's analysis, as applicable to the present experimental pattern, continues as follows (1960, Sec. 6, P. 12):

> There are 70 ways of choosing a group of 4 objects out of 8. This may be demonstrated by an argument familiar to students of "permutations and combinations", namely, that if we were to choose the 4 objects in succession we should have successively 8, 7, 6, 5 objects to choose from, and could make our succession of choices in 8 x 7 x 6 x 5, or 1680 ways. But in doing this we have not only chosen every possible set of 4, but every possible set in every possible order; and since 4 objects

> can be arranged in order in 4 x 3 x 2 x 1, or 24 ways, we may find the number of possible choices by dividing 1680 by 24. The result, 70, is essential to our interpretation of the experiment.

Fisher notes that the subject could, at best, judge rightly in each one of the eight instances, and since there are four of each kind, points out (1960, Sec. 6, pp. 12-13) that

> ....this amounts to choosing, out of the 70 sets of 4 which might be chosen, that particular one which is correct. A subject without any faculty of discrimination would in fact divide the 8...correctly into two sets of 4 in one trial out of 70, or, more properly, with a frequency which would approach 1 in 70 more and more nearly the more often the test were repeated. Evidently this frequency, with which unfailing success would be achieved by a person lacking altogether the faculty under test, is calculable from the number of [films] used. The odds could be made much higher by enlarging the experiment, while, if the experiment were much smaller even the greatest possible success would give odds so low that the result might, with considerable probability, be ascribed to chance.

(The implications of this last consideration will be dealt with later in more detail in considering the sensitiveness and precision of experiments; see Chapters 8 and 11.)

For the purpose of gaining a clear understanding of fundamental considerations in the design of experiments, a particular advantage of this experiment (shared also in the experiment subsequently described on pp. 24 ff. ) is that the basic probability argument involved in the interpretation of the results can be followed in full by most readers, *without having to take anything on faith*. This contrasts with the situation found in the case of most statistical tests of "significance" (or, to use the term we prefer, statistical *confidence* tests), in which most researchers, unless they are mathematicians, must indeed take a good deal on faith. That is, they use a formula, and look up the resulting value in a published table. Even if (as is frequently not the case) the basic logic of the formula is at least dimly understood, its precise derivation and the assumptions underlying it generally are not. And even in this rare instance, the tabled values ordinarily cannot be checked and must be accepted as published, with only the assurance that *someone* (presumably someone wiser than the reader and/or presumably infallible) *has* checked and found them correct.

Now, we are not in any way trying to question the accuracy of statistical tables, most of which have, in fact, been so carefully compiled and so thoroughly checked and rechecked as to be of almost unvarying accuracy (though mistakes do occur once in a great while). Rather, the reason for preferring to present at least one or two examples which the reader need *not* take on faith is that we strongly feel (admittedly without proof to back up our feeling) that a complete understanding of the logic, and complete following of its implications through to a specific conclusion, is *per se* of great value in the education of an experimental investigator. (We feel that it is also an esthetic experience.) The primary attitude to which we hope doing this will conduce is to see that there *is* a rigorous and comprehensible logic underlying statistical tests; and the tendency which we hope it fosters is to examine this logic as fully as possible as it applies in a specific investigation. Only this tendency, carried as far as the training of the investigator and his consultant advisors can permit, will, we believe, lead to the avoidance of basic errors and misconceptions in the design of experiments. The need to follow completely the logic of inquiry is eloquently stated in the second of the two quotations given by Fisher as prologue to *The Design of Experiments*. Taking a realistic rather than idealistic viewpoint of the linguistic attainments of many of our readers, we offer the following rough translation of this statement in the present context :

> The only means of forestalling these errors consists of suppressing, or at least of simplifying as much as possible, our reasoning, which is *ours*, and which can alone mislead us; to put it continually to the test of *experience,* not to cleave to anything except the *facts*, which are the truths given us by nature and which alone cannot mislead us; not to look for truth except in the series of experiences and observations--above all, in the order in which they are presented--in the same manner that mathematicians arrive at the solution of a problem by the simple arrangement of givens and by *reducing reasoning to operations so simple, to judgments so direct, as never to lose sight of the evidence which must serve as their guide.*

[Author's translation from the original French, as quoted by R. A. Fisher, of A. L. Lavoisier's *Methode de Nomenclature Chemique* [1787)].

In the present illustrative experiment, in the one to be presented next, and in all instructional media experiments to be subsequently considered, this need for this close reasoning--"*à des opérations si* *simples, a des jugemens si courts, qu'ils ne perdent jamais de vue l'évidence qui leur sert de guide*"--will apply both in the calculation of the probabilities that are essential to a meaningful interpretation of the experimental results and, as well, in the argument by which the operations actually performed in conducting the experiment are related to these calculations in such a way that the interpretation can, indeed, be shown to be a rigorously definable one.

### The test of significance and its sensitiveness

Fisher's remarks here (1960, Sec. 7, pp. 13-14) are also well worth quoting in full (with the bracketed expressions again substituted in accordance with the details of the variant parallel example we are using in this presentation):

> *It is open to the experimenter to be more or less exacting in respect to the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him.* Thus, if he wishes to ignore results having probabilities as high as 1 in 20--the probabilities being of course reckoned from the hypothesis that the phenomenon to be demonstrated is in fact absent--then it would be useless for him to experiment with only 3...[films] of each kind. For 3 objects can be chosen out of 6 in only 20 ways, and therefore complete success in the test would be achieved without sensory discrimination, i.e., by "pure chance," in an average of 5 trials out of 100. It is usual and convenient for experimenters to take 5 per cent, as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach

> this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. *No such selection can eliminate the whole of the possible effects of chance coincidence*, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 trials is decidedly "significant," in the statistical sense, we thereby admit that *no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon*; for the "one chance in a million" will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*. *In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure*. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

In the case of our illustrative experiment, we may decide that if every film is rightly classified, a "significant" or satisfactorily trustworthy positive conclusion as to the film producer's discriminative ability should be made. In other words, we will agree that we should admit that in this event he had made good his claim. Now, however, let us ask what our conclusion should be if, for each kind of film his judgments are *three right and one wrong*? Given the conditions as adopted, and presuming that these are kept in mind by our subject, we may assume that any error in one set of judgments will be compensated by an error in the other (since it is known to the subject that there are four films of each kind.) As Fisher points out in connection with the parallel tea-tasting experiment (1960, Sec. 7, p. 14):

> In enumerating the number of ways of choosing 4 things out out 8, such that 3 are right and 1 wrong, we may note that the 3 right may be chosen, out of the 4 available, in 4 ways and, independently of this choice, that the 1 wrong may be chosen, out of the 4 available, also in 4 ways. So that in all we could make a selection of the kind supposed in 16 different ways. A similar argument shows that, in each kind of judgment, 2 may be right and 2 wrong in 36 ways, 1 right and 3 wrong in 16 ways and none rights and 4 wrong in 1 way only. It shold be noted that the frequencies of these five possible results of the experiment make up together, as it is obvious they should, the 70 cases out of 70.
>
> It is obvious, too, that 3 successes to 1 failure, although showing a bias, or deviation, in the right direction, could not be judged as statistically significant evidence of a real sensory discrimination. For its frequency of chance occurrence is 16 in 70, or more than 20 per cent. Moreover, it is not the best possible result, and in judging of its significance we must take account not only of its own frequency, but also of the frequency of any better result. In the present instance "3 right and 1 wrong" occurs 16 times, and "4 right" occurs once in 70 trials, making 17 cases out of 70 as good as or better than that observed. The reason for including cases better than that observed becomes obvious in considering what our conclusions would have been had the case of 3 right and 1 wrong only 1 chance, and the case of 4 right 16 chances of occurrence out of 70. The rare case of 3 right and 1 wrong could not be judged significant merely because it was rare, seeing that a higher degree of success would frequently have been scored by mere chance.

The "null" hypothesis

The term "null hypothesis" has been previsouly encountered (see Chapter 3, ) and enters, implicitly or explicitly, into every experimental hypothesis test. The role of the null hypothesis in the present illustrative experiment is described by Fisher as follows (1960, Sec. 8):

> Our examination of the possible results of the experiment has therefore led us to a statistical test of significance, by which these results are divided into two classes with opposed interpretations. Tests of significance are of many different kinds, which need not be considered here. Here we are only concerned with the fact that the easy calculation in permutations which we encountered, and which gave us our test of signficance, stands for something present in every possible experimental arrangment; or, at least, for something required in its interpretation. The two classes of results which are distinguished by our test of significance are, on the one hand, those which show a significant discrepancy from a certain hypothesis [namely, in our example, the hypothesis that the judgments given are in no way influenced by the way in which the films were reproduced]; and on the other hand, results which show no significant discrepancy from this hypothesis. This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation. Much confusion would often be avoided if it were explicitly formulated when the experiment is designed. In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said

> to exist only in order to give the facts a chance of disproving the null hypothesis.[4]

## Randomization

The mathematical logic of the model followed by the present illustrative experiment should, we believe, be transparently clear to any intelligent person, without mathematical training, from a careful study of the foregoing. We now have to consider what *physical operations* in the conduct of the experiment are necessary so that this mathematical model can indeed be defended as appropriate. It is at this point that the importance of randomization, previously stressed, can be appreciated as what Fisher calls "the physical basis for the validity of the test." (1960, Sec. 9)

> We have spoken of the experiment as testing a certain null hypothesis, namely, in this case, that the subject possesses no sensory discrimination whatever of the kind claimed; we have, too, assigned as appropriate to this hypothesis a certain frequency distribution of occurrences, based on the equal frequency of the 70 possible ways of assigning objects to two classes of 4 each; in other words, the frequency distribution appropriate to a classification by pure chance. We have now to examine the physical conditions of the experimental technique needed to justify the assumption that, *if discrimination of the kind under test is absent, the result of the experiment will be wholly governed by the laws of chance.*

As Fisher points out, this might well be otherwise. For example, in our illustration, all of the reversal films might be silent and all of the negative films accompanied by sound; or, all of the former might have been projected on the lefthand side of a screen, by one projector, with all of the latter projected at the right, and/or with a different projector; or all of one kind might have been used many times and badly scratched in the process, whereas those of the other kind were fresh, new prints. Any of these differences could well result in all of those similarly processed being classed alike. These sets might then be classified either all right or all wrong; but, in either event, the frequency of the critical occurrence in which all the films were to be classified *correctly* would no longer be *1* in 70, but rather *35* in 70. As Fisher points out, though the test of significance would then be wholly vitiated, "errors equivalent in principle to this are very frequently incorporated in otherwise well-designed experiments."

A crucial further point is that it is no help to insist that all the films (or whatever else we may be dealing with) must be "exactly alike" in every particular except the one that is to be tested. For, as Fisher points out, *"this is a totally impossible requirement in our example, and equally in all other forms of experimentation."* In our illustration, for example, it is likely that the point of view from which the films were photographed, or the illumination of the objects photographed will differ slightly; or that the illumination provided by the projector may vary perceptibly from the first to the last film showing, etc. That these are only examples of differences that might be present is obvious; what may be less obvious is (Fisher, 1960, Sec. 9) that it:

> ...would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable. When

any such cause is named, it is usually perceived that, by increased labour and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment. Our view, which will be much more fully exemplified in later sections, is that it is an essential characteristic of experimentation that it is carried out with limited resources, and an essential part of the subject of experimental design to ascertain how these should be best applied; or, in particular, to which causes of disturbance care should be given, and which ought to be deliberately ignored. To ascertain, too, for those which are not to be ignored, to what extent it is worth while to take the trouble to diminish their magnitude. For our present purpose, however, it is only necessary to recognise that, whatever degree of care and experimental skill is expended in equalising the conditions, other than the one under test, which are liable to affect the result, this equalisation must always be to a greater or less extent incomplete, and in many important practical cases will certainly be grossly defective. We are concerned, therefore, that this inequality, whether it be great or small, shall not impugn the exactitude of the frequency distribution, on the basis of which the result of the experiment is to be appraised.

## The effectiveness of randomization

The account given next by Fisher in respect to "The Effectiveness of Randomization" is a classic in cogency and clarity. As Fisher (1960, Sec. 10 points out, the

> ...element in the experimental procedure which contains the "essential safeguard" is that the [two films are to be presented]..."in random order." This, in fact is the only point in the experimental procedure in which the laws

of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced. The phrase "random order" itself, however, must be regarded as an incomplete instruction, standing as a kind of shorthand symbol for the full procedure of randomisation, by which the validity of of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated. To demonstrate that, with satisfactory randomisation, its validity is, indeed wholly unimpaired, let us imagine all causes of disturbance...to be predetermined...then since these, on the null hypothesis, are the only causes influencing classification, we may say that the probabilities of each of the 70 possible choices or classifications which the subject can make also are predetermined. If now, after the disturbing causes are fixed, we assign, strictly at random, 4 out of 8... [films] to each of our experimental treatments, then every set of 4, whatever its probability of being so classified will certainly have a probability of exactly 1 in 70 of being the 4...[for example, the four films developed by the reversal process]. However important the causes of disturbance may be, even if they were to make it certain that one particular set of 4 should receive this classification, the probability should be the same, must be rigorously in accordance with our tests of significance.

It is apparent, therefore, that the random choice of the objects to be treated in different ways would be a complete guarantee of the validity of the test of significance, if these treatments were the last in time of the stages in the physical history of the objects which might affect their experimental reaction.

The fact that experimental treatments can not, in practice, always be applied last, is not necessarily a cause for concern, since subsequent causes of differentiation, provided they are under the experimenter's control (for example, the choice of different projectors or classrooms to be used with different films in an instructional experiment) can either be predetermined before the treatments have been randomized (so that they are then automatically randomized also), or, if this is not done, can later be independently randomized on their own account. Thus (Fisher, 1960, Sec. 10)

> ...other causes of differentiation will be either (a) consequences of differences already randomised or (b) natural consequences of the difference in treatment to be tested, of which on the null hypothesis there will be none, by definition, or (c) effects supervening by chance independently from the treatments applied.

Therefore, aside from any error resulting from the experimenter's introducing, through lack of proper procedure, other extraneous differences in treatment along with test treatments (or subsequently)...

> ...the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged.

Sensitiveness of the experiment

An objection which the subject or the reader might make to the experiment as thus far described is that only if every film is classified correctly will the subject be judged successful; a single error will reduce his performance below the agreed-upon level of confidence. The subject might well contend, not that he could make the discrimination with invariable correctness, but merely that, though sometimes mistaken, he would be right most of the time, or at least would be more often than not. It is now apparent that, in order to test this less extreme claim, the experiment must either be enlarged, or in some other way made more sensitive to detecting deviations from the null hypothesis less extreme than those considered previously (i.e., all judgments correct).

The way in which the required size of the experiment might be arrived at, if we make no change in procedure, can be seen by extending Fisher's previous calculations on which the statistical test for the eight-film experiment was shown to depend. For example, the reader may verify that an experiment with 12 films, (six of each kind) would predict, on the null hypothesis, one chance out of 924 for five out of the six of each kind to be classified right and one out of the six to be classified wrong.

Thus with a 12-film experiment, our subject's claim could potentially be substantiated at some level of confidence if he judged five or more times correctly, and there are 37 ways (36 plus 1) in which this could happen. Since 37 is less than a twentieth of 924, such a test would merit the agreed-on standard of confidence, although one pair of films may have been wrongly classified; and it can be easily verified that, using still larger numbers of films, the desired degree of confidence could be obtained with a still higher proportion of errors. Thus by increasing the size of the experiment, we can always render it more sensitive, meaning that we can thus provide for the confident demonstration of a lower degree of discrimination--or, in other words, a quantitatively smaller departure from the null hypothesis. And, since in every case a satisfactory experiment is capable of disproving the null hypothesis, the sensitiveness of an experiment is increased whenever it permits the null hypothesis to be disproved through the presence of a less extreme difference or smaller degree of discrimination.

There are various ways in which the sensitiveness of any experiment might be increased--generally, however, at the expense of some increase of cost or trouble. One way that is always available, at least in principle, is to make the experiment larger. In the present illustrative experimental pattern (Fisher, 1960, Sec. 11):

> The same result could be achieved by repeating the experiment, as originally designed, upon a number of different occasions, counting as a success all those occasions on which [all 8] are correctly classified. The chance of success on each occasion being 1 in 70, a simple application of the theory of probability shows that 2 or more successes in 10 trials would occur, by chance, with a frequency below the standard chosen for testing significance; so that the sensory discrimination would be demonstrated, although, in 8 attempts out of 10, the subject made one or more mistakes. This procedure may be regarded as merely a second way of enlarging the experiment and, thereby, increasing its sensitiveness, since in our final calculation we take account of the aggregate of the entire series of results, whether successful or unsuccessful. It would clearly be illegitimate, and would rob our calculation of its basis, if the unsuccessful results were not all brought into the account.

## Other methods of increasing sensitiveness

Instead of enlarging or replicating the experiment, we may also attempt to increase its sensitiveness in other ways. The first of these is through use of a different experimental structure. Adapting Fisher's discourse to our present example, we can illustrate a change of structure if, instead of determining in advance that four films (chosen at random) should be of each kind, we might instead determine independently, by chance, the treatment of each successive film to be seen by our subject,--e.g., by the toss of a coin--so that each treatment has an equal chance (i.e., a probability of exactly one-half) of being chosen each time. The probability of classifying correctly eight such events by pure chance is evidently 1 in $2^8$, or 1 in 256 chances.* Also, there are only 8 chances out of 256 of classifying seven correctly and one wrong (see Appendix B ). This being the case, the sensitiveness of the experiment could be increased with this mode or pattern of experimental design, though still using only 8 films; it would be possible to demonstrate "significant" discrimination even if one film were classified incorrectly. Thus, as Fisher points out (1969, Sec. 12):

> In many types of experiment, therefore, the suggested change in structure would be evidently advantageous. For the special requirements of a psychophysical experiment, however, we should probably prefer to forego this advantage [since it would occasionally occur that all 8 of the films would be of one kind] and this, besides bewildering the subject by an unexpected occurrence, would deny [him] the real advantage of judging by comparison.

Experimental arrangements in which the assignment to treatments of the members of each of a number of pairs is determined by chance are frequently applicable in instructional-media research; however, an example will be given presently.

## Refinements of technique or experimental procedure

The other general way in which an experiment ought to be made more sensitive is through certain refinements in technique. These, as Fisher emphasizes, "contribute nothing to the validity of the experiment, and of the test of significance by which we determine its results. They may, however, be important, and even essential in permitting the phenomenon under test to manifest itself." The use of "refinements of technique" to increase the sensitiveness of instructional-media experiments is further discussed in Chapter 11 . Here however, we may well stress the role of such refinements in relation to the sensitiveness and validity of experiments through the words of Fisher (1960, Sec. 12), who says:

> How far such experimental refinements should be carried is entirely a matter of judgment, based on experience. The validity of the experiment is not affected by them. Their sole purpose is to increase its sensitiveness, and this object can usually be achieved in many other ways, and particularly by increasing the size of the experiment. If therefore, it is decided that the sensitiveness of the experiment should be increased, the experimenter has the choice between different methods of obtaining equivalent results; and will be wise to choose whichever method is easiest to him, irrespective of the fact that previous experimenters may have tried, and recommended as very important, or even essential, various ingenious and troublesome precautions.

## The place of stimulus discrimination in the study of instructional media

It may be noted, finally, that the problem of discrimination as here presented deals with the discriminative ability of one person; a more general question would ask to what extent the same discriminative ability can be exhibited by others. Particular interest would attach to the question of whether the discrimination revealed for the film producer would also be found in those for whom, if for anyone, the discrimination would be likely to make a conceivable difference in our real "payoff" interest--viz., the effectiveness of film-media instruction. These would evidently not be the producer of the films, but rather the students who are to be instructed by the films.

Two other points are also worth making. First, the ability to discriminate on the basis of graininess in the film might represent one which, while demonstrable when a given judge is attending solely to this aspect of the film, would nevertheless not be noticed under more ordinary circumstances--i.e., when attention was on learning what the film was presenting rather than on the more artificial task of discriminating photographic-reproduction "quality". Second, even if the discrimination was evidenced in the latter situation, and even if the difference was apparent to an audience of school learners as well as to the producer, it would by no means necessarily follow that the perception of the photographic-quality difference would necessarily carry with it a difference in amount learned from the film--which of course, is our concern in the present work. (For some experimental data on the effect of filmic "quality" differences, see May and Lumsdaine (1958) and Zuckerman (1953) summarized in Chapter 14, p. ).

Finally, it should be re-emphasized that all of the general principles here exemplified, as well as many of the more specific considerations of method, would apply in any instructional-media experiment in which any kind of effect of a program of any kind was under investigation.

## EXPERIMENT II: The Effect of an Auto-Instructional Program in Spelling on Knowledge of the Subject Matter

Let us now consider a second illustrative experiment in which the same basic principles, already exemplified in the discrimination experiment based on Fisher's description, also apply. Here we shall again stress the central role of probability calculations and hopefully shall also convince the reader that in the case of a simple but widely applicable experimental structure he can, even though his mathematical training might have stopped short in high school, follow the full logic of the experimental proof without having to take on faith any mysterious formulas or published tables of unknown derivation.

This experiment is again a hypothetical one; we have so chosen it not only for ease but also, we hope, for clarity of exposition. By ignoring the usual procedural complications for the present and concentrating on the essential logical structure, we hope to avoid the distraction of procedural details and foster understanding of the essentials. In due course, the application of this same experimental model will be shown in several real instructional-media experiments for which the data as actually obtained will be examined, and in which other important aspects of the whole of experimental design will be considered.

## The crucial role of probability or "betting odds" in experiments

As has been seen in the preceding example, the notion of an experiment (in the scientific sense) always involves the concept of probability. It may be possible to by-pass the explicit consideration of probability in simple experiments in the physical sciences (though often, even then, at some risk of being fooled), but in any experimentation that involves the effects of something on human behavior (which is of course true of all experiments on the effects of educational procedures or media), probability must enter explicitly.

*The variability of behavior*. The reason why this must be true can be expressed in several ways. Perhaps it is sufficient here to remind the reader that behavior (e.g., achievement on an achievement test) characteristically is subject to many influences, and that any given index of human performance almost always shows wide variation (a) from person to person and (b) from situation to situation. The way in which these facts lead us to have to consider the application of probability concepts in an experiment is readily seen by considering a few extreme examples. The cautions against being misled by failure to reckon with probability in these extreme cases apply with equal force, even if not always so obviously, in less extreme instances.

*What are the chances that "A" knows more than "B" (or "F" than "G," etc.)*? For example, suppose we experimentally give a spelling program to one person, Mr. F, and then test his knowledge of spelling, as compared with the knowledge of one other person, Mr. G, who did not have the program. It is obvious that we cannot conclude much from the results of such a comparison. If after taking the program Mr. F knows more about spelling than does Mr. G, we must inevitably ask *what are the chances* that he also knew more about spelling *before* he took the program. If we had in fact given the program to neither individual, it is almost certain that one or the other would know somewhat more (from past experience) than the other. Let us further suppose that the comparison was unbiased as far as deciding which person was to get the program (say, we decided it by the toss of a coin), but that we have no further information about the two individuals. Then what are the chances that, *without* having taken the program, F knew more than G? Fifty-fifty, we will say. Put otherwise, there is an even chance that F, being chosen at random, would score higher than G (we will assume a detailed enough test to ignore the chances of their getting exactly equal scores) *regardless of whether* he took the program or not. Hence it clearly would be folly, without further information, to conclude just from F's higher score after taking the program that the higher score was an "effect" or "result" of taking the program. That F's score after taking the program is higher than G's (when G didn't take it) thus tells us *nothing* dependable about the effect of the program.

*Proving that a program had an effect*. How then can we prove to the satisfaction of even a mildly skeptical questioner (all those who would base decisions on evidence--not just scientists--must be at least this skeptical) that the program did in fact have some effect on knowledge? The way this can be done (and in principle the *only* way it can be done in any non-trivial case) is to invoke the notion of probability. Suppose, for example, that we have not just one Mr. F but *100* F's, and also 100 G's, instead of just one Mr. G. Further, let us suppose that the decision as to which of an available 200 persons were to take the program and which not was again determined by chance (i.e., randomly). Then if *all 100* Fs score higher than any of the Gs, we may well conclude that, whatever else the reasons for this result may include, the result is probably *not* just a chance occurrence. Why? Because, of course, it is intuitively obvious that the probability that *all 100* Fs (when chosen at random) would know more initially than a randomly chosen 100 Gs is so low as to be negligible for most purposes. Unlike the case with only one F and one G, it is, we may say, now so improbable that we will easily risk the chances of conceivably being wrong in concluding that the difference between F and G was a mere rare chance fluke. That is, we will conclude (provided we further know that treatment of Gs and Fs differed, after being randomly chosen, only in whether or not they were given the program) that the program had a positive effect on knowledge of spelling. (The importance of that one parenthetical proviso will be returned to later on.)

Randomization and probability. Now the concept of probability has entered into even so simple a comparison as this in two ways: First, in interpreting the results, we thought inevitably in terms of the likelihood that, just by chance, one of our two individuals (in the first instance), or each person in two sets of 100 individuals each (in the second instance) might show knowledge superior to his "opposite number." In the first instance we concluded that this probability was too high to be ignored, and in the second instance we concluded, even without any explicit arithmetical calculation, that it was low enough to be ignored. But this conclusion in terms of probability was self-evident only because of another way in which we introduced probability--in our assignment of persons to be Fs or Gs at random, or by chance. It was this that enabled us to proceed from the necessary assumption of equal probability that one or the other of the two was superior before they took the program. Only by making some such use of the notion of equal probability, or lack of bias, in the initial assignment was our later conclusion about the effects of the program made defensible. To take another extreme possibility, if we had assigned 100 college seniors to be the Fs and 100 high school freshmen to be the Gs, we might well expect the Fs to be greatly superior before they took the program. Hence our finding them consistently superior after the program would be a very doubtful basis for concluding that the difference in knowledge was the result of the program. Our identifying it as such is most clearly and simply evident only if there is in fact an equal chance that any one individual is to be chosen to be an F or a G.

Simplicity of essential experimental logic. Now in these simple and extreme examples lies much of what is basic to the validity of experiments in educational media, and yet it is surprising, when we consider situations less obvious (because less extreme or less simply described) how often the basic need for appropriate consideration of probability is overlooked in interpreting data that purport to indicate media's effects. Much of the methodology with which this volume deals is in fact concerned with how observations about differences between groups with different exposure to media can be interpreted validly in terms of basic notions of probability when the situation is less obvious or clear cut than those just described. In this methodology, so essential to the forming of valid conclusions about the effects of media, we shall have to deal repeatedly--even though we stick to relatively simple basic experimental designs--with explicit considerations of probability and the so-called laws of chance as they operate in what are sometimes called "tests of significance." In doing so, we believe, however, the essential logic of the argument can be followed without elaborate discussion of probability concepts beyond what is evident from relatively simple, if less extreme, comparisons such as those just made.

Pitfalls in the calculation of odds

However, the calculation of probabilities is full of snares unless certain fundamentals are kept clearly in mind. It can also get very complex, rapidly getting us into advanced mathematics in many situations, including some which at first sight look deceptively simple (as the reader can readily verify by consulting such a text as that of Uspensky's).* But adherence to a few fundamental principles will suffice to reduce to easy straightforward arithmetical calculation a wide variety of problems, including the statistical test used in the first experiment described above, and in the further one concerning the determination of a program's effects, to which we shall return shortly.

Two such basic principles will be reviewed in a moment. To see the snares that can be involved if certain of them are not clearly apprehended, we may first consider a simple-sounding gambling game.

This illustration of the deceptive quality of probability occurs in Leslie Charteris' short story, "The Percentage Player."* In the story, the problem is presented as a gambling game, "Dong Hai," that affords an easy way to fool most of the people all of the time--at a profit. It also provides an interesting exercise in logic, and points out how easy it is to make a mistake in calculating odds--simply by not properly taking stock of all the possibilities.

The game involves three small discs or poker chips: one with an "X" (or other mark) on both sides, the second so marked on only one side, and the third with both faces blank. The player conceals the chips from his sight, by dropping them in a bag or in some similar fashion, and draws one at random, not knowing which it is. He looks at only one face of the chip and attempts to predict what he will see when he turns the chip over--that is, he tries to predict whether or not the unexposed side is marked with the X.

The question raised in Charteris' story is whether, given an X showing on the chip, it is possible to tell with any better than chance likelihood what the other side will be. Is there an even chance of an X or a blank on the other side, or is there a better than fifty-fifty chance for one or the other?

The answer seems obvious; as one of the story's characters remarked, the game is a fancy substitute for flipping a coin. After all, one reasons, if we see an X, it must be on one of the first two chips: the double-X or the X-blank. Therefore, one further reasons, the other side is either an X or a blank; and, since we don't know which it is, the best one can do is thus to guess, and the odds dictate that the guess will be right about fifty percent of the time. Thinking thus of the problem as a gambling game, there would thus seem to be no system of winning consistently over a period of any length.

Almost anyone will reach this conclusion (which is quite erroneous) after a brief examination of the facts of the problem--because almost nobody will initially take all of the equally likely possibilities into account. Yet this answer is wrong, and any gambler who has not been misled by appearance will make money playing the game. But every possibility must be considered to arrive at the correct answer.

What really are all the possibilities if an X is drawn? To answer correctly, we must ask not just which chip the X could be marked on but which X, which individual face, could be showing. How many X marks are there, in all? Three, of course; on each side of the double-X chip and on one side of the X-blank. Therefore, the X, which has been randomly drawn, could be any of three different marks; knowing nothing more, we must consider each as an equally likely possibility.

If we have drawn the X-blank chip, naturally the back of our X shows a blank. However, if we are faced with one side of the double-X, the other side is also an X. And if we are faced with the third possible mark, the second side of the double-X, then the hidden side is again marked. Therefore, in two out of three possible cases with an X showing, the other side is also an X.

The same is true if a blank face is drawn. In one case, the X-blank chip, the other side is marked; but if either of the other two blanks (on the double-blank) is showing, the hidden side is a blank. What, then, can we actually say about the odds? We can say that, having drawn a chip, its hidden face will be the same as the side that is showing more often than not (on the average, two out of three times): If we draw an X, the chances are two to one (i.e., two out of three) that the other side is also an X; if we draw a blank, the chances are also two out of three the other side will also be blank. The gambler, then, will win, on the average, two out of three times, if he simply "guesses" that the two sides of the chip he has drawn are the same.

The basic error made in the initial erroneous calculation of probabilities, which is almost always made on first impulse in this game, is failure to include all possible combinations in the reckoning. For our purposes, it will suffice to keep in mind the definition of probability of success (or failure) of any event as the ratio of all the number of possible, independent and equally likely ways in which an event can turn out to be successful (or, alternatively, unsuccessful if we are interested in the probability of failure) to the total number of all possible, independent, equally likely outcomes.*

### Probability of compound and disjunctive events

The basic principles of probability with which we need here be concerned include, in addition to the basic definition given above, those which govern the probability of multiple or alternative outcomes. The argument underlying these is familiar to many readers (being presented in many elementary algebra texts), but it will be useful to state the two principles here:

(1) Compound events. The probability that all of several independent outcomes will eventuate is the product of the independent probabilities. For example, if we assume that the probability of getting a head in tossing a coin is one half, the probability of both heads in two independent tosses is 1/2 x 1/2, or 1/4; for 3 heads out of 3 tosses is 1/2 x 1/2 x 1/2, or 1/8, etc. The probability that tossing seven coins will yield all seven heads in any one instance is $(1/2)^7$ or 1/128.

(2) Disjunctive events. The probability that either of two mutually exclusive events will eventuate is the sum of their separate probabilities. (If the events are, furthermore, together exhaustive, the sum of their probabilities will total to 1.00.)

For example, since the probability of getting any one number from one to six in a toss of a single die is 1/6 (if the die is fair), the pro ability of getting either of two specified faces (say, either a five or a six) is 1/3; the probability of getting a one or a two or a three is 1/2;

the probability of not getting a given number (say, a one) is 5/6--that is, one minus the chances of getting a two or a three or a four or a five or a six. Similarly, the probability of getting either a head or a tail in tossing a theoretically fair coin is 1/2 plus 1/2 or 1.00.

It is important to note that probabilities are thus additive only if the events concerned are mutually exclusive. When we consider various situations involving outcomes which can be achieved in more than one way, this is often not the case. For example, when we inquire about the chances of getting a given face--say an ace--on either of two tosses of a die (or on one of two dice tossed simultaneously), a naive person will often suppose the probability is 1/3. The absurdity to which this leads is evident, however, as soon as we consider what we should say of the probability of getting the ace in six tosses or seven tosses. Clearly in six tosses the probability (P) is not 1.00, as we would be led to conclude by continuing to add the probabilities of 1/6 per toss for the successive tosses. Obviously, we might frequently fail to get an ace in any of the six tosses. (Even more absurd, if possible, is the extrapolation to seven or more tosses, where the probability would, by the same additive-probability reasoning, exceed 1.00!)*

The trouble, of course, is that the events are not mutually exclusive--e.g., we might get an ace on more than one toss. The moral is, once again, that valid calculation of probabilities--whether in games of chance or in statistical tests for assessing the reliability of experimental data--must always consider all of the possibilities, and must take care to insure that assumptions about equal likelihood are really justified. The fallacy of assuming equal chances for an X or a blank in the poker-chip game involved the violation of both of these cardinal principles in concert.

## Calculation of probabilities in determining the effect of a program

Now let us return to the problem of determining whether an auto-instructional program (or other instructional-media program) has an effect where we have better evidence than merely that from the score made by one Mr. F and one Mr. G, but we have less extreme evidence than consistent superiority of all the scores for 100 Fs over those for any of the 100 Gs. Faced with evidence less extreme than this consistent superiority of every person in the 100-to-100 comparison, how can we still arrive at dependable conclusions on which reasonable men can agree? The single reference experiment described below is intended to show all the essentials of how this can be done.

In the first place, lacking results as obviously undependable as our previous one-versus-one man comparison or, on the other hand, as obviously "non-chance" as our previous 100-vs.-100 comparison, we must be able in some fashion to calculate the probability of obtaining, by chance, particular outcomes less extreme than these. In addition, we need some way of doing one of two things: Either (1) we must agree on some uniform way to report the value of probability (P) arrived at, leaving it to the reader what conclusion to draw from this value; or (2) we must agree on some probability value which we will regard as a cut-off point, below which we can agree to dismiss the results as non-chance and above which we will regard effects as unproven. (Note that we do not say "above which we will say there are no effects," but merely that we will regard the evidence as insufficient to show that there are effects: The non-equivalence of these two statements has already been stressed by Fisher [see pp. 12-13], and will be the subject of considerable later discussion.)

The educational value of a statistical test of "transparent" clarity

The more statistically sophisticated researcher knows of many ways to calculate probabilities. However, he may not always understand the basic logic involved--and small wonder (unless he is a mathematician of some competence going beyond ordinary integral calculus)--since the derivation of theorems on which are based most of the significance tests that generate these probabilities call for considerable mathematical attainments even to follow them--much less to master them. In addition, in the more complex forms of analysis the basic underlying logic is often not directly displayed in the sometimes relatively complicated calculations that generate the end result, especially when, for the sake of efficiency, these calculations are reduced to a set of cookbook rules and formulas in which the final comparison is many steps removed from the basic data.

A simple reference model that displays the basic logical steps by which a probability value can be derived from data has already been provided in the illustration, based on Fisher's account, in the first part of the chapter. A second example is presented here in the hope that the basic logic that underlies it can be simply reviewed not only to the benefit of the neophyte but perhaps also to the profit and refreshment of the statistics-weary researcher of more experience. The latter may have computed endless series of t-tests, analyses of variance or even, alas, co-variance--but with only incomplete initial understanding of this basic logic, or with what he once understood so overwoven by computational fabric that he has all but forgotten the essentially beautiful simplicity of the underlying logical framework.

Indeed, it is a rare researcher who ever really understood all the steps involved in a t-test; therefore most researchers must take on faith the correctness of the table in which they look up the final result because of never having checked how it was derived. (These derivations of the proof for the unbiased estimate of variance that is involved require advanced calculus; and while no student of algebra would presumably verify all the entries in his logarithm table, it is much more satisfying to use the table if one at least knows the steps whereby the entries could in principle be verified if one had a mind to do so.) The advantage of the test presented below is that while widely applicable to diverse experimental situations, it also satisfies the criterion expressed by the previously cited quotation from Lavoisier, presented by Fisher as prologue to *The Design of Experiments*. The crux of this statement, particularly relevant here to the presentation of a "transparent" probability test, may bear repetition: "...*reducing reasoning to steps so simple, to judgments so direct, as never to lose sight of the evidence on which they are based.*"

While the test to be presented is less sensitive (but sometimes not greatly so) than more elaborate tests based on such hidden derivations and mysterious or taken-on-faith tabulations, it has the advantage that *all* of its steps from initial assumptions to final conclusion can be followed with no more mathematics than the simplest algebra. Many students seem to find it peculiarly satisfying and refreshing to the faith that other more complex methods could, similarly, be found to be embedded in the bedrock of unassailibly clear reasoning, should they have the time to follow them out. This test, the binomial or sign test, is, furthermore, often a fairly effective test which can be quickly and easily applied in assessing the tenability of hypotheses in many everyday situations as well as in at least preliminary analysis of a great deal of the kind of data that are likely to be generated by educational-media experiments. Finally, it can be readily reconstructed in terms of a simple mnemonic device by the French mathematician, Pascal; once learned, the test could, if need be, be recalled (for use even on a deserted island!) with no reference books available.

We present an example of how this might be done, in the context of a reference experiment. The same test can be applied also to several other reference experiments described later, in which the "workings" for more complex tests used are less fully transparent, less readily reconstructed, or less universally applicable.

This simplest of all tests involves only a slight departure from the basic "F" versus "G" comparisons suggested above. It involves, first of all, the notion of pairs of observations on "C" and on "E" (which we will substitute, for variety, for the previous symbols "F" and "G").

## Binomial test for a "C versus E" experiment in spelling

For our hypothetical experiment let us now assume that we have an indefinite number of pairs of individuals, and that we can assign at random one member of any pair to a "C" treatment and the other member to an "E" treatment. We might suppose that the E condition is the working through an experimental auto-instructional program and C, the Control condition the withholding of this program, or that C and E are two alternative program treatments, or whatever we wish. For the sake of specificity let us say that treatment E is the giving of one form of self-instructional program on spelling, which we hypothesize on the basis of a theory to be a superior method, and that C is an alternative form of spelling instruction (perhaps a different form of program with larger frames or less cueing). In any case, the particular intent of the program, the respect in which the treatments differ and the kind of outcome to be assessed (the nature of the dependent variable) are all irrelevant for our present purpose.

With respect to the dependent variable, we might take as an example the scores made by individuals, following each treatment, on a spelling test; for the present, however, we shall simply assume that the outcome is to be decided by a spelling "bee," in which the E student will be pitted against the C student for each pair of students; and that the fact of whether the winner of each pair was in the E group or the C group will be taken as one observation favoring the E program if the E student wins and favoring the C program if the C student wins.

The basic logic of the statistical test now required is this: Of any pair of observations assigned at random to two conditions, C and E, we should, without any prior information, expect by chance that C or E would be superior (eliminating ties) just half the time in the long run. Thus, we expect indefinitely repeated tosses of a coin to come out half heads and half tails. This is intuitively obvious if we define a fair coin as one that has an equal chance of coming up either way, though it is equally obvious to any one with much coin-tossing experience that in any short series of tosses the actual frequency of heads will depart somewhat from exactly one half of the total tosses.*

That a long series of tosses would in fact converge closer and closer to .50000... cannot actually be proved rigorously without more extensive mathematical reasoning.** But we believe that the practically minded researcher will accept intuitively the notion of equal probability based on any of several random methods of assignment for the members of any one pair. If so, we can proceed with a satisfying and wholly "transparent" test by continuing to assign independently one member of each pair to C and one to

E. We thus wind up with N pairs, each of which would be equally likely, if in fact the performance of individuals following the C and E treatments differed only by chance, to contain an E member who would beat the C member.

Now, having administered a treatment to the E members which differs from that administered to the C members, we compare the results of a test given to each E member and each C member of each pair, asking only which member won the spell-down over his "opposite number." Let us suppose that more E members won--that is, more E's were superior to the corresponding C members than vice versa. We now ask: What are the chances that this result could reasonably have arisen by chance--i.e., the probability of obtaining the result if only chance were operating. (The fact probably is that we do not wish, in fact, to accept the hypothesis that only chance is at work. This hypothesis--that only chance is at work--is our "null hypothesis" and is a possible explanation of any set of results that we want to rule out, at some reasonable level of confidence, before accepting the conclusion that the difference in experimental treatment--rather than mere chance--was responsible for any observed after-the-treatment difference between E and C.

Obviously, the same would also apply if our paired observations were a set of test scores, or a set of paired before and after scores for a set of individuals: We would want to ask whether a given frequency of instances in which there was improvement (after-score better than before-score) could be dismissed as quite likely to occur just as a result of chance variation in an individual's performance on two different occasions.

Calculation for perfectly consistent results. First, let us take the situation in which all E members win out over their opposite members within each pair, and examine several cases for varying numbers of pairs. Within only one pair, the result that E is better gives us no confidence in for him

rejecting the null hypothesis, since the probability is .50 to come out better just by chance. Let us then see what level of confidence in showing a non-chance cause-effect relation could be obtained by having *two* pairs instead of just one. Since the outcome of each pair is *independent*, we have four *equally likely outcomes* for the two pairs, viz.:

| | | | | |
|---|---|---|---|---|
| Pair 1: | E | C | E | C |
| Pair 2: | E | E | C | C |

One of these four equally likely outcomes—only one—shows C superior for both pairs, two show C and E superior once apiece, and one shows E superior for both pairs. Thus we should expect that on the average we will find E superior for both pairs (or two times out of two), once in every four times by chance alone. Otherwise stated, the chances of an E, E outcome are one in four, or the odds are thus three to one against this outcome if only chance is at work.

This still is not good enough odds to satisfy usual criteria for rejecting chance on a plausible explanation, so we proceed to consider *three* pairs. Here there are eight ($2^3$) possible outcomes, all equally likely:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Pair 1: | E | E | E | C | E | C | C | C |
| Pair 2: | E | C | E | E | C | E | C | C |
| Pair 3: | E | E | C | E | C | C | E | C |
| | (3 Es) | (2 Es) | | | (1 E) | | | (0 Es) |

The probability of *all three* Es occurring by chance is thus now only one in eight (.125). Without enumerating, we can see that with four pairs there are $2^4$ or sixteen possible outcomes so that the chances of getting *all* Es are only one in sixteen (.0625); with five pairs there are $2^5$ possible outcomes and the probability that E would be superior by chance on all five is one in thirty-two (.031+), and so on, so that generally the chances of *all* Es with N pairs are 1 in $2^N$.

Now we may demand any level of confidence (e.g., degree of smallness for the probability of getting a result) that we please; see the previous comment by Fisher, and also the discussion of rejection level in Chapter 5. If we want to accept the influence of a non-chance causal factor (e.g., the hypothesized superiority of our "E" program) as having been demonstrated whenever the probability of getting a given result by chance is less than one in ten, it is evident that four pairs will suffice (thus giving P of .0625) *if we get all four Es superior*. If, as is more customary, we insist on a probability (P) at least as small as $P=.05$, we shall, with perfectly consistent results, at a minimum need five pairs in order to reject the null hypothesis of only chance differences. If we insist on $P=.01$, as is sometimes done, then we must have at least seven pairs (for six pairs, $2^N=64$, or $P=1/64$; only with seven pairs [or more], when $P=1/2^N=1/128$, do we go below the 1/100 level). For the rest of this discussion, let us consider that about .05 is an adequately low level of probability—that is, that at this level of probability or lower we would prefer to reject the null hypothesis and conclude that some non-chance factor (presumably, the difference in experimental treatment represented by the two spelling programs) is demonstrated (or "proved" to our satisfaction, at least) to be operating.

*Calculation when results show an exception.* We have seen that if five (or more) pairs consistently favor E *without exception*, this criterion is met. Now, however, what if the results favor E in most but not all cases?

To determine the probability of any successful outcome we have to consider two numbers, a numerator, which is the number of equally likely ways in which a successful outcome can occur, and a denominator, which is the total number of outcomes, each equally likely, that could occur. When we deal with outcomes in which there are no failures (all pairs are successful) as above, there is only one way (e.g., E, E, E, for three pairs) in which this successful outcome can occur, and the numerator is thus always unity ($\frac{1}{2^N}$).

The situation is only slightly more complex when we consider the equally likely ways in which there is only one failure (i.e., all but one of the pairs favor E). As seen in the above enumeration there are three ways this can happen with three pairs, so that the chances of getting two or more Es in three pairs are thus 4 in 8, or fifty-fifty. With four pairs, we can obviously get all-but-one in four ways, namely EEEC, EECE, ECEE, or CEEE; that is, the one failure can be any one of the four pairs. Thus with four pairs the chances of getting three out of four "successes" are 1/4, and the chances of getting three out of four or better (i.e., either 3 out of 4 or all 4 out of 4) are 1/4 + 1/16, or 5 in 16. Obviously, this can be extended; in general, with N pairs there are N ways of getting one failure--i.e., a success (E better than C) in all-but-one cases--since the one failure can be in any of the N pairs. Since we have seen that the total number of possible outcomes is $2^N$, the probability of getting all-but-one success is $\frac{N}{2^N}$. Since the chance of getting no failures is $\frac{1}{2^N}$, the probability of all-but-one-or-fewer is therefore $\frac{1}{2^N}$ plus $\frac{N}{2^N}$, or $\frac{N +}{2^N}$ for five pairs, this gives us 6/32, for six pairs 7/64, for seven pairs 8/128, for eight pairs 9/256, and so on. Since 8/128 = .069 approximately, and 9/256 = .035 approximately, we evidently need at least eight pairs to reject the null hypothesis at the agreed on criterion of having P less than .05, when one failure will be allowed as well as perfectly consistent successes.

How about proving our case for a non-chance effect (i.e., rejecting or disproving the null hypothesis that there are only chance effects) when there are even less consistent results--say all-but-two successes, all-but-three, etc.? To cope with such possibilities we need only extend our calculations keeping in mind the basic principle of determining the total number of equally likely ways in which each event could occur. For example, with six pairs (for which the total possible number of outcomes is $2^6$ or 64), in how many ways could we obtain all but two successes? This amounts to asking in how many ways two failures could be chosen out of six possible opportunities to fail, using an argument similar to that employed in Fisher's account of "permutations and combinations" in the previous experiment (Chapter 4, p. 5). Since one of the failures could be in any of the six pairs, and the other one in any of the remaining five, there are a total of 6 x 5 or 30 possible ways in which the two failures could be chosen. But half of these are duplicates (e.g., fifth pair with third pair as well as third pair with fifth pair), differing only in the order in which they are named. Thus there are only 30 ÷ 2 or 15 different combinations.

Similarly, for 8 pairs, we can obtain 2 failures in (8 x 7) ÷ 2, or 28 ways, for 9 pairs in (9 x 8) ÷ 2, or 36 ways, for 10 pairs in (10 x 9) ÷ 2 ways, and so on. How about three failures? In the case of 8 pairs, for example, it is evident, by the same kind of argument, that these can be chosen in $8 \cdot 7 \cdot 6$ ways, or permutations, but that of the 3 members of each such trio (e.g., the second, fourth and fifth pairs) there will be undesired duplications, differing only in the order in which they are named. Specifically, the first named of each trio can be any of the 3 in that particular combination of pairs, the second named can be any of the remaining two, and the last will have to be the remaining one. There are thus 3·2·1 or 6 arrangments for each possible set of three pairs, and, since we have just seen that the total number of ways of choosing the 3 failures is 8·7·6, the number of different combinations of pairs which can be the 3 failures is $\frac{8\cdot7\cdot6}{3\cdot2\cdot1}$ or 56.

Similarly, the number of ways in which we could get four failures out of 8 pairs is: $\frac{8\cdot7\cdot6\cdot5}{4\cdot3\cdot2\cdot1}$, or 70. (Evidently the number of ways in which we could get 4 successes out of 8 is given by exactly the same calculation, as was seen in Fisher's argument as applied to Experiment I.)

More generally, in determining the number of ways in which we could get k failures out of N pairs, we can see that the number of ways in which the k failures can occur is given by the expression $\frac{N!}{k!\,(N-k)!}$, where the "!" following any number "X" (X!, read "X factorial") means the product $X(X-1)(X-2)\cdots3\cdot2\cdot1$. This is apparent if we note that, as exemplified in our example for 3 failures in 8 pairs, the value $\frac{8\cdot7\cdot6}{3\cdot2\cdot1}$ can also be written as $\frac{8\cdot7\cdot6\cdot5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)\,(5\cdot4\cdot3\cdot2\cdot1)}$, that is: $\frac{8!}{3!\,(8-3)!}$, with the last five factors (5·4·3·2·1, or 5!) in the numerator cancelling out the same five factors in the denominator. Similarly for four failures, we have:

$$\frac{8!}{4!\,(8-4)!}=\frac{8!}{4!\,4!}=\frac{8\cdot7\cdot6\cdot5\cdot\cancel{4}\cdot\cancel{3}\cdot\cancel{2}\cdot\cancel{1}}{(4\cdot3\cdot2\cdot1)\,(\cancel{4}\cdot\cancel{3}\cdot\cancel{2}\cdot\cancel{1})}=\frac{8\cdot7\cdot6\cdot5}{4\cdot3\cdot2\cdot1}=70.$$

In the case of ten pairs we would get:

0 failures (all 10 successes) in only- - - - - - - - - - 1 way

1 failure in any of the 10 pairs, or - - - - - - - - - -10 ways

2 failures in $\frac{10\cdot 9}{2\cdot 1}$, that is, in- - - - - - - - - - - - 45 ways

3 failures in $\frac{10\cdot 9\cdot 8}{3\cdot 2\cdot 1}$, that is, in- - - - - - - - - - 120 ways

4 failures in $\frac{10\cdot 9\cdot 8\cdot 7}{4\cdot 3\cdot 2\cdot 1}$, that is, in - - - - - - - - 210 ways

5 failures in $\frac{10\cdot 9\cdot 8\cdot 7\cdot 6}{5\cdot 4\cdot 3\cdot 2\cdot 1}$, that is, in- - - - - - - - 252 ways

For 6, 7, 8, 9, and 10 failures we can simply extend the numerator and denominator, obtaining successively 210, 120, 45, 10 and 1, or if we are intelligently lazy, we can observe that the series must be a symmetrical one [e.g., since the number of ways for 2 successes (or 8 failures) is again 45]. The total series: 1, 10, 45, 120, 210, 210, 120, 45, 10, 1, will, as a check, sum to 1024, that is, $2^{10}$, as of course it must since there are $2^{10}$ possible outcomes in all. Finally, we may note that the probability of getting two or fewer failures will be $(1 + 10 + 45) \div 1024$, or 56/1024, or about .055, which thus falls just short of our self-adopted "conventional" criterion of .05 or less--to fully meet which we should therefore need (with as many as 2 failures) more than 10 pairs.

Now let us go back, however, to some smaller numbers of pairs and review the application of the argument that has been presented. Starting out with just one pair we can, merely, either succeed once or fail once--that is, we have 1 way of failing and one of succeeding, out of 2 possible ways. For 2 pairs: 1 way of getting no failure, 2 ways of getting 1 failure and 1 way of getting no successes (both pairs fail)—out of a total of $2^2$ or 4 ways. For 3 pairs:

1 way to get 0 failures,

3 ways to get 1 failure,

3 ways to get 2 failures, and

1 way to get 3 failures,

out of a total of $2^3$ possible ways. Let us tabulate this as follows:

| number of pairs (N) | ways to get k failures, for k = 0 | 1 | 2 | 3 | total number of possible outcomes = $(2^N)$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | | | 2 |
| 2 | 1 | 2 | 1 | | 4 |
| 3 | 1 | 3 | 3 | 1 | 8 |

Applying the combinational principles developed above for determining the number of ways in which any given number of failures could occur in N pairs will, as the reader can now easily verify, allow us to extend this tabulation as follows:

| N: No. of pairs | k: number of failures 0 | 1 | 2 | 3 | 4 | 5 | 6 | . . . | total ways $(2^N)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | | | | | | | 2 |
| 2 | 1 | 2 | 1 | | | | | | 4 |
| 3 | 1 | 3 | 3 | 1 | | | | | 8 |
| 4 | 1 | 4 | 6 | 4 | 1 | | | | 16 |
| 5 | 1 | 5 | 10 | 10 | 5 | 1 | | | 32 |
| 6 | 1 | 6 | 15 | 20 | 15 | 6 | 1 | | 64 |
| . | | | | | | | | | . |
| . | | | | | | | | | . |
| . | | | | | | | | | . |

Note that each row adds across to $2^N$, the total number of possible outcomes for that number of pairs.

Now, note also that the body of this table shows, on examination, an interesting fact which when seen affords a ready mnemonic for reconstructing the table, or extending it, without going through the arithmetic labor of computing for each entry the number of ways in which k failures could occur in N pairs, or instances, by the formula $\frac{N!}{k!(N-k)!}$. This is the fact that starting with line 2, for 2 pairs, each and every entry in the table is exactly the sum of the number immediately above it plus the number, if any, just to the left of that number. For example: for 3 pairs, 1 = 1 + 0, 3 = 2 + 1, 3 = 1 + 2, and 1 = 0 + 1; or, for 5 pairs, 1 = 1 + 0, 5 = 4 + 1, 10 = 6 + 4, 10 = 4 + 6, 5 = 1 + 4, and 1 = 0 + 1.

If the body of the table is written in a slightly different way we have, for the first four lines,

1 1
1 2 1
1 3 3 1
1 4 6 4 1

which is called Pascal's triangle, after its inventor, the French mathematician Blaise Pascal, and in which we can readily see that the next three lines will be:

1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1

each succeeding line being obtained by adding the two adjacent numbers in the line immediately above. The proof that this will be true in the general case involves only a relatively simple algebraic argument, the essential basis for which the reader may possibly have encountered in his high school algebra book (see Appendix B); the upshot is that we can generate the lines of the table giving the number of ways that 0, 1, 2, 3, . . ., k, . . . N − 1, or N failures could occur out of N independent pairs (or N independent instances, more generally, if the probability of failures is 1/2 for each instance) by simply adding on lines in Pascal's triangle up to any number N that we want.

Either in the form of Pascal's triangle or in the tabulation with the straightened columns, as in the second tabulation on page 4-46, the additive method of generating the successive lines is, once practiced two or three times, easily remembered, and together with the check of adding each line out to $2^N$, provides an easy basis for even the "deserted island" reconstruction alluded to above.

It may also be noted (as is obvious on reflection) that the chances for N - k failures (or N - k successes) are equal to those for k failures (or k successes), that is, the triangle is symmetrical, as it must be since the probability for k successes must equal that for k failures. Also note that since for any number of pairs, N, there will be N + 1 entries (including 0 failures as well as 1, 2, 3, . . ., N failures), and there will be an even number of "mirror image" entries when N is odd (e.g., six entries: 0, 1, 2, 3, 4, 5, for N = 5), but an odd number of entries (with one central unique value) when N is even. Finally, a very interesting further fact is that the successive numbers entered across any line of the table, for N pairs are, in fact, the numerical coefficients of the binomial expansion $(a + b)^N$ — or, for that matter $(x + y)^N$, or anything-plus-anything to the Nth power—for example, $(a + b)^2 = a^2 + 2\,ab + b^2$, with coefficients 1, 2, 1; $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$; etc. (See Appendix B for the general formula and its relation to the above-discussed factorial quotient for the number of ways of getting k failures out of N pairs.) It is for this reason that the test of confidence in experimental data, the argument underlying which we have been tracing here in hopefully short, transparently clear steps, is called the binomial test--or, as a slight variant, the sign test [since it considers for each pair, as in our spelling "bee" only the direction or sign (plus or minus) of the outcome rather than considering quantitative differences in spelling-test "scores" or similar quantitative measures].

Let us now complete this argument, reverting temporarily to consideration of just the cases up to 4 pairs. First, we can express the probability of getting k failures in our series of N (1, 2, 3, or 4) pairs or trials, as a set of fractions, by taking the number of equally likely ways in which we can get k failures as the numerator of each fraction and the total number of equally likely outcomes as its denominator. This gives us:

| N | k (number of failures) | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| 1 pair (or 1 instance): | 1/2 | (1/2) | | | |
| 2 pairs (or 2 instances): | 1/4 | 2/4 | (1/4) | | |
| 3 pairs (or 3 instances): | 1/8 | 3/8 | (3/8) | (1/8) | |
| 4 pairs (or 4 instances): | 1/16 | 4/16 | 6/16 | (4/16) | (1/16) |
| which also are the probabilities for: | all | all but 1 | all but 2 | all but 3 | all but 4 successes. |

However, in view of the symmetry of these distributions, all we need at most is the probability values in the left-hand position (or "one tail," as it is called) of such a table, thus:

| | k = 0 | 1 | 2 |
|---|---|---|---|
| N = 1 | 1/2 | | |
| N = 2 | 1/4 | 2/4 | |
| N = 3 | 1/8 | 3/8 | |
| N = 4 | 1/16 | 4/16 | 6/16 |

(or)

| | k = 0 | 1 | 2 |
|---|---|---|---|
| N = 1 | 1/2 | | |
| N = 2 | 1/4 | 1/2 | |
| N = 3 | 1/8 | 3/8 | |
| N = 4 | 1/16 | 1/4 | 3/8 |

[Note here that the additive mnemonic rule--one above plus one to its left for the Pascal-triangle kind of tabulation does not apply to the fractional probabilities but only to the numerator terms (for fractions in which the denominator is $2^N$).] Since probability values are generally expressed in decimal values we might translate the above as in the following tabulation:

| | k = 0 | 1 | 2 |
|---|---|---|---|
| N = 1 | .5000 | | |
| N = 2 | .2500 | .5000 | |
| N = 3 | .1250 | .3750 | |
| N = 4 | .0625 | .2500 | .3750 |

Before completing the calculation of such a table for N>4, however, let us observe that, as pointed out by Fisher (p. 11 ) we are always interested in the chances of getting *k* failures (or *k* successes) *or fewer.* Therefore, corresponding to the partial (left-hand) tabulation for the probabilities we might as well translate the table to give the sums of the number of ways of getting *k or fewer* failures, thus keeping to integral numbers as long as possible before doing any dividing. Thus for N = 1, 2, 3, and 4, corresponding to the tabulation for exactly *k* failures--namely:

| | k = 0 | 1 | 2 | 3 | 4 | $2^N$ |
|---|---|---|---|---|---|---|
| N = 1 | 1 | 1 | | | | 2 |
| N = 2 | 1 | 2 | 1 | | | 4 |
| N = 3 | 1 | 3 | 3 | 1 | | 6 |
| N = 4 | 1 | 4 | 6 | 4 | 1 | 8 |

we have, for k *or fewer* failures:

| | k ≤ 0 | 1 | 2 | 3 | 4 | $2^N$ |
|---|---|---|---|---|---|---|
| N = 1 | 1 | 2 | | | | 2 |
| N = 2 | 1 | 3 | 4 | | | 4 |
| N = 3 | 1 | 4 | 7 | 8 | | 8 |
| N = 4 | 1 | 5 | 11 | 15 | 16 | 16 |

Here the last term of the cumulative frequencies in each row of course equals the entry in the $2^N$ column, since all of the possibilities have been cumulated by that point.

- In Table 1-A (page 4-52 ) we present the numbers for the frequencies (number of ways of getting) exactly k failures for N up through 10, continuing the series for each row through the first "repeat" (marked by parentheses) in the symmetrical series, so that the reader can, if he wishes, verify the numbers a column at a time, starting with the left-hand column of ones, by application of the Pascal rule. Table 1-B (page 4-52 ) then cumulates these frequencies up to and including (but to save unnecessary addition, not beyond) the point in each row after which the frequencies in Table 1-A start to repeat.

Before dividing the entries by $2^N$ for each row to get a table of probabilities we can save some further work, if we are interested only in probability values of around .05 or less as a basis for rejecting a null hypothesis of chance difference, by inspecting the table to see where this region lies. Thus, it is apparent that only the values that have been underlined in Table 1-B are smaller than one-tenth of $2^N$ for their row, so that probabilities as small as 0.1 will be obtained only for these values. Dividing each such entry by $2^N$ and re-orienting the table 90° for convenience, gives us the following probabilities of less than 0.1 for the chances of getting k or fewer failures--that is, getting all but k, or fewer, successes--out of N independent trails (e.g., randomly assigned pairs) if only chance is operating:

Probability for all but k (or fewer) successes

| k | N . . . 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| 0 | 1/16 | 1/32 | 1/64 | 1/128 | 1/256 | 1/512 | 1/1024 |
| 1 | | | | 8/128 | 9/256 | 10/512 | 11/1024 |
| 2 | | | | | | 46/512 | 56/1024 |

**Table 1-A**

| No. of pairs | Relative frequency (number of ways of getting) exactly k (= 0, 1,...etc.) failures<br>0 | 1 | 2 | 3 | 4 | 5 | 6 | Total ways =($2^N$) |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | (1) | | | | | | 2 |
| 2 | 1 | 2 | (1) | | | | | 4 |
| 3 | 1 | 3 | (3) | | | | | 8 |
| 4 | 1 | 4 | 6 | (4) | | | | 16 |
| 5 | 1 | 5 | 10 | (10) | | | | 32 |
| 6 | 1 | 6 | 15 | 20 | (15) | | | 64 |
| 7 | 1 | 7 | 21 | 35 | (35) | | | 128 |
| 8 | 1 | 8 | 28 | 56 | 70 | (56) | | 256 |
| 9 | 1 | 9 | 36 | 84 | 126 | (126) | | 512 |
| 10 | 1 | 10 | 45 | 120 | 210 | 252 | (210) | 1024 |

**Table 1-B**

| No. of pairs | Relative frequency of getting k or fewer failures (k ≦ 0, 1, 2,....,etc.)<br>0 | 1 | 2 | 3 | 4 | 5 | $2^N$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | 2 |
| 2 | 1 | 3 | | | | | 4 |
| 3 | 1 | 4 | | | | | 8 |
| 4 | 1 | 5 | 11 | | | | 16 |
| 5 | I | 6 | 16 | | | | 32 |
| 6 | I | 7 | 22 | 42 | | | 64 |
| 7 | I | 8 | 29 | 64 | | | 128 |
| 8 | I | 9 | 37 | 93 | 163 | | 256 |
| 9 | I | 10 | 46 | 130 | 256 | | 512 |
| 10 | I | 11 | 56 | 176 | 386 | 638 | 1024 |

It has already been noted that unless we have highly consistent results (2 or fewer failures out of even 10 pairs) we shall need somewhat more pairs than 10. On the other hand, the reader can guess that for any great degree of consistent superiority in an experiment such as that concerning our spelling programs (or any two other experimental treatments) we shall need much fewer than the 100 pairs suggested in our original extreme example (p. 26 ).

On the other hand, were we so confident of the superiority of one of our experimental programs as to expect perfectly consistent results, it is evident that more than 7 or 8 pairs would hardly be useful since we would probably draw the same conclusion (i.e., reject the null hypothesis) with the probability value of 1/1024 or 1/512 as with a slightly larger value such as 1/128 (which is less than .01) or perhaps even with the P-value of 1/64 which we could obtain from only six pairs.

An extension of the tabulation of probabilities, for N of up to 50, converted into decimal form is given in Table 2 (pp. 4-53A, B) for the regions of interest where P ranges between around 0.1--higher values than which (indicated in the table by "+") normally are not considered as being of significance as evidence for rejecting a null hypothesis--down to around .01--lower values than which (indicated in the table by "-") are normally not interpreted much differently from .01 so far as the conclusions to be drawn are concerned.* (The reader should by now be able to verify this table for himself as far as his skepticism and industry may dictate.) A graphical plot of "critical" probability values (i.e., those likely to be adopted as criteria of acceptance or rejection), based on Table 2, yields Figure 1 (p. 4-53C), in which the ordinate shows the number of failures that can be tolerated, for the abscissa values of N up to 50,

| k \ N= | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | .5000 | .2500 | .1250 | .0625 | .031 | .016 | .0078 | .0039 | .0020 | .0010 |
| 1 |  | + | + | .31 | .19 | .11 | .0625 | .035 | .020 | .011 |
| 2 |  |  |  | + | + | + | .23 | .14 | .090 | .055 |
| 3 |  |  |  |  |  | + | + | + | .25 | .17 |
| 4 |  |  |  |  |  |  |  | + | + | + |

| k \ N= | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | - | - | - | - | - | - | - | - | - | - |
| 1 | .0059 | .0032 | .0017 | .0009 | - | - | - | - | - | - |
| 2 | .033 | .019 | .011 | .0065 | .0037 | .0021 | .0012 | .0007 | - | - |
| 3 | .11 | .073 | .046 | .029 | .018 | .011 | .0064 | .0038 | .0022 | .0013 |
| 4 | + | .19 | .13 | .090 | .059 | .038 | .025 | .015 | .0096 | .0059 |
| 5 | + | + | + | .21 | .15 | .11 | .072 | .048 | .032 | .021 |
| 6 |  | + | + | + | + | + | .17 | .12 | .084 | .058 |
| 7 |  |  |  | + | + | + | + | + | .18 | .13 |
| 8 |  |  |  |  |  | + | + | + | + | + |

| k \ N= | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | - | - | - | - | - | - | - | - | - | - |
| 3 | .0007 | - | - | - | - | - | - | - | - | - |
| 4 | .0036 | .0022 | .0013 | .0008 | - | - | - | - | - | - |
| 5 | .013 | .0085 | .0053 | .0033 | .0020 | .0012 | .0008 | - | - | - |
| 6 | .039 | .026 | .017 | .011 | .0073 | .0047 | .0030 | .0019 | .0012 | .0007 |
| 7 | .095 | .067 | .047 | .032 | .022 | .014 | .0096 | .0063 | .0041 | .0026 |
| 8 | .19 | .14 | .11 | .076 | .054 | .038 | .026 | .018 | .012 | .0081 |
| 9 | + | + | + | .15 | .11 | .084 | .061 | .044 | .031 | .021 |
| 10 | + | + | + | + | + | .16 | .12 | .092 | .068 | .049 |
| 11 |  | + | + | + | + | + | + | .17 | .13 | .010 |
| 12 |  |  |  | + | + | + | + | + | + | + |

Table 2 (part 1)

Probability of k or fewer failures out of N independent instances (when p for each instance = 1/2)

| N=<br>k | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | - | - | - | - | - | - | - | - | - | - |
| 7 | .0017 | .0011 | .0007 | - | - | - | - | - | - | - |
| 8 | .0053 | .0035 | .0023 | .0015 | .0009 | - | - | - | - | - |
| 9 | .015 | .010 | .0068 | .0045 | .0030 | .0020 | .0013 | .0008 | - | - |
| 10 | .035 | .025 | .018 | .012 | .0083 | .0057 | .0038 | .0025 | .0017 | .0011 |
| 11 | .075 | .055 | .040 | .029 | .020 | .014 | .010 | .0069 | .0047 | .0032 |
| 12 | .14 | .11 | .081 | .061 | .045 | .033 | .024 | .017 | .012 | .0083 |
| 13 | + | + | .15 | .11 | .088 | .066 | .049 | .036 | .027 | .019 |
| 14 | + | + | + | + | .16 | .12 | .094 | .072 | .054 | .040 |
| 15 | + | + | + | + | + | + | .16 | .13 | .10 | .077 |
| 16 |  | + | + | + | + | + | + | + | .17 | .13 |
| 17 |  |  |  | + | + | + | + | + | + | + |

| N=<br>k | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|
| 9 | - | - | - | - | - | - | - | - | - | - |
| 10 | .0007 | - | - | - | - | - | - | - | - | - |
| 11 | .0022 | .0014 | .0010 | - | - | - | - | - | - | - |
| 12 | .0058 | .0040 | .0027 | .0018 | .0012 | .0008 | - | - | - | - |
| 13 | .014 | .0098 | .0069 | .0048 | .0033 | .0023 | .0015 | .0010 | .0007 | - |
| 14 | .030 | .022 | .016 | .011 | .0080 | .0057 | .0040 | .0028 | .0019 | .0013 |
| 15 | .059 | .044 | .033 | .024 | .018 | .013 | .0093 | .0066 | .0047 | .0033 |
| 16 | .11 | .082 | .063 | .048 | .036 | .027 | .020 | .015 | .011 | .0077 |
| 17 | + | .14 | .11 | .087 | .068 | .052 | .039 | .030 | .022 | .016 |
| 18 | + | + | + | .15 | .12 | .092 | .072 | .056 | .043 | .032 |
| 19 | + | + | + | + | + | .15 | .12 | .097 | .076 | .059 |
| 20 | + | + | + | + | + | + | + | .16 | .13 | .10 |
| 21 |  | + | + | + | + | + | + | + | + | .16 |
| 22 |  |  |  | + | + | + | + | + | + | + |

Table 2 (part 2)

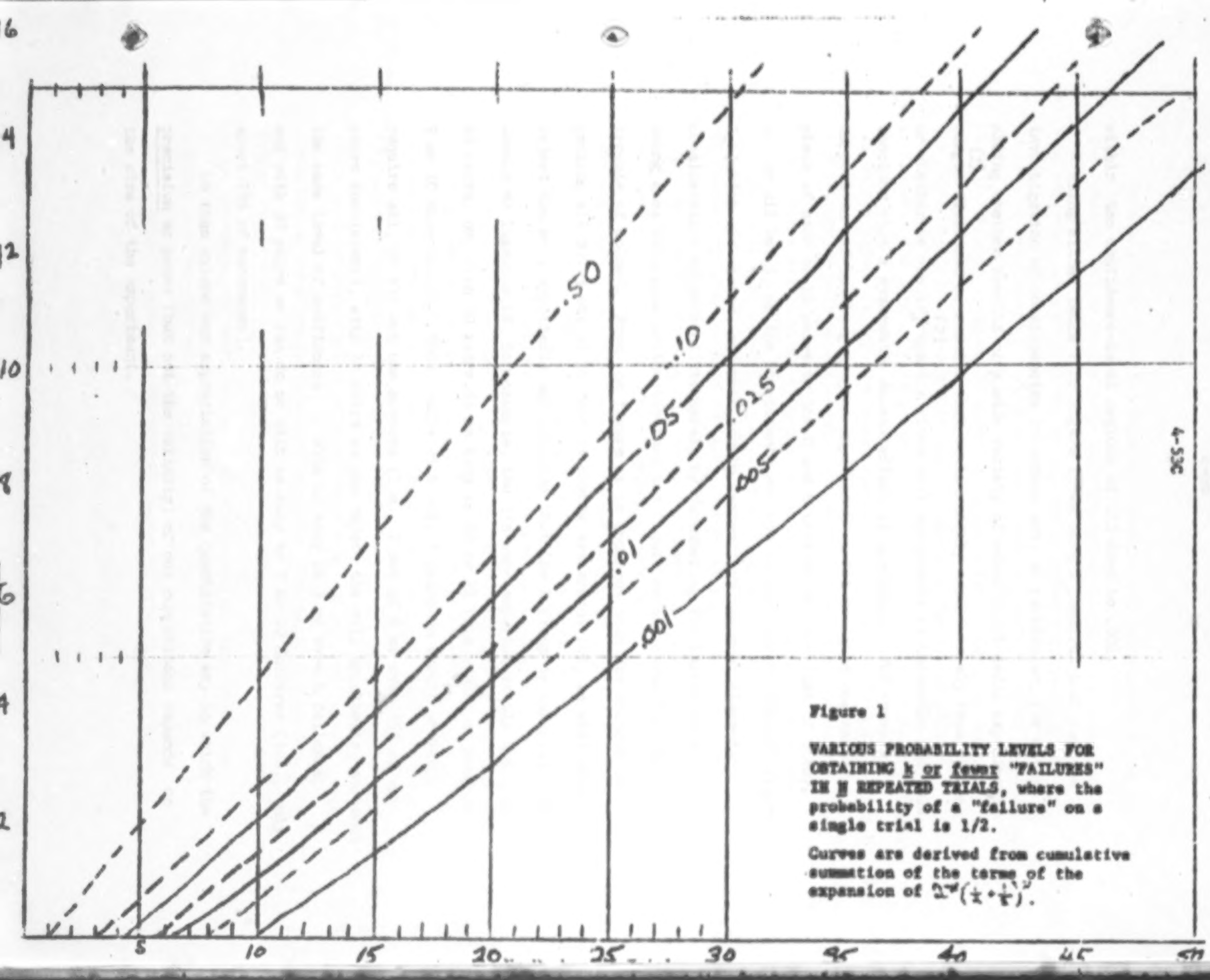


Figure 1

**VARIOUS PROBABILITY LEVELS FOR OBTAINING k or fewer "FAILURES" IN N REPEATED TRIALS**, where the probability of a "failure" on a single trial is 1/2.

Curves are derived from cumulative summation of the terms of the expansion of $2^{-N}\left(\frac{1}{2}+\frac{1}{2}\right)^{N}$.

within the confidence-level regions of .10 down to .001.

Using either Table 2 or Figure 1, we have a powerful tool for the investigation of consistencies in nature and, in particular, for determining whether results of a wide variety of educational-media experiments can (1) be dismissed as results that would readily result merely from chance, or whether we should (2) regard a given null hypothesis as untenable and thus conclude that a systematic causal effect is operating. For example, suppose in our spelling-program experiment we had assigned one member apiece of each of 16 pairs to the C and E treatments, and that of these, 13, or all but 3, of the 16 students won out in the "bee." Consult Figure 1, (which the reader, remember, could derive from the most elementary consideration of probability merely by remembering the Pascal rule and doing some laborious arithmetic that the figure merely saves him the trouble of doing). From the figure it is apparent that the chances of getting all but 3 out of 16 just by chance are around .01; we will thus reject the null hypothesis and conclude that the E group is superior. We should do likewise if, for example, the E's were superior in all but 4 of 20 pairs; or, with 30 pairs, in as many as 20 or 21 (i.e., with as many as 9 or 10 exceptions). Thus, while with only 8 pairs we should probably require all, or all but one success (i.e., 7 out of 8 or over 85% of our pairs consistent), with 16 pairs we can reject the null hypothesis, at about the same level of confidence with as many as 3 or even 4 failures, and with 30 pairs we can do so with as many as 9 or 10 failures (i.e., only about 70% of successes).

We thus extend our appreciation of the quantitative way in which the precision or power (but not the validity) of our comparisons depends on the size of the experiment.

We shall return to this kind of consideration in Chapters 8 and 11. For the present, let us merely note that the matter can also be looked at the other way around: that is, we may use an advance estimate of the degree of consistency which our confidence in the potency of an experimental variable leads us to expect as a basis for planning an experiment so as to determine the probable size of the experiment that will be required. The kind of test we have developed here will be quite satisfactory if we can expect fairly consistent results. With less basis for expected consistency, a more sensitive kind of test will be preferable--one which will utilise not only the direction of superiority but also the amount.

We shall consider briefly such statistical tests in the latter part of Chapter 8, after we have dealt a bit more systematically with the question of measurement as applied to dependent variables in an experiment. The basic point which has been illustrated here is the dependence of interpretation of an experiment's outcomes on probability calculations whose nature must, for valid interpretation, derive squarely from the kind of operations performed in the experiment. As noted earlier, the advantage of the statistical test we have derived for use in Experiment II is that all of the steps in its logic can be fully traced out even by the mathematical novice. A similar logic, but one requiring more involved mathematical reasoning, more arduous calculation, or both, underlies all valid tests of significance of confidence that can be used for the interpretation of experimental data. In applying such tests the reader, with proper statistical consultation may proceed with confidence if he remembers the essential logical considerations from which the present test was derived.

In order to attain an economical degree of precision with a relatively small sized experiment, and where it is important to detect less potent effects than would be possible with this simpler binomial test, statistical tests less directly derivable (for the average reader) are generally appropriate. Student's t test and some forms of the analysis of variance are among the leading examples, together with such rank-ordering tests as Wilcoxon's.* However, even when such tests are needed, the reader often can still find practical use for the simply derivable binomial test in making a quick, rough assay of the dependability of the results. This may be useful and satisfactory on first inspection of new data from an experiment, or on checking between results presented in others' research reports. And not infrequently, this first rough test --with its dual advantage of "transparent" logic clearly understood and its care of application--will lead to essentially the same conclusion as would eventually be arrived at by a somewhat more complex, less obvious, and less fully understood test.*

"Two-tailed" statistical tests

It will be noted that the probability calculations and interpretations we have made from the binomial or sign test** have been based on either exceeding a given number of failures or a given number of successes, but not both. That is, we have reduced the chances of getting as many as k failures, out of N instances, for E to be superior to C. Any given degree of consistency--say, 13 out of 16, or only 3 failures in 16 pairs or instances--is logically accompanied by the opposite possibility of getting, in this case, only 3 successes (all but 3 failures) out of the 16 instances.

It may be helpful here to present the complete probability distribution for 16 pairs, which can readily be derived from Pascal's triangle (or from the coefficients of a binomial expansion for the 16th power ). This gives, out of a total of $2^{16}$ or 65, 536, the following relative frequencies:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Failures | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Successes | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 |
| Relative frequency | 1 | 16 | 120 | 560 | 1820 | 4368 | 8008 | 11,440 | 12,870 |

(cont.)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Failures | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Successes | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
| Relative frequency | 11,440 | 8008 | 4368 | 1820 | 360 | 120 | 16 | 1 |

The probability for each outcome is of course merely the indicated frequency divided by 65,536. It may be helpful to visualize this frequency distribution in a graphical plot, in which relative frequency is plotted as the vertical (ordinate) distance for each of the 17 possible categories of number of successes or failures. Unless such a plot is made on a very large scale, the frequencies for the more extreme outcomes are too low to be readily distinguished visually from zero, as shown in Figure 2, p. 4-57A.

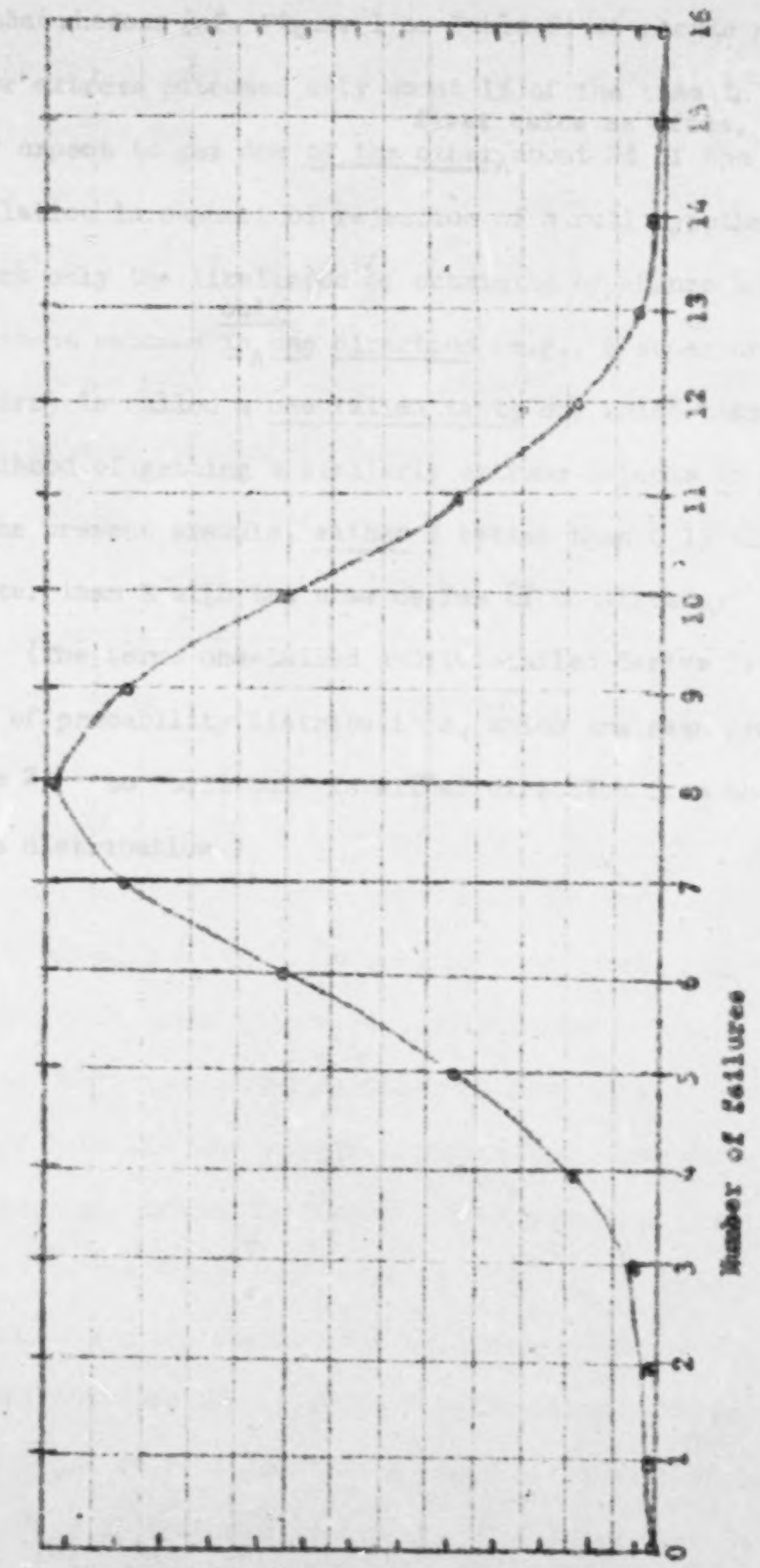

Relative Frequency
13,000
12,000
11,000
10,000
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of Failures

Figure 2

In view of the symmetry of the binomial distribution, it is easy to see that whereas (cf. Figure 1 or Table 2) we should get either one of these rather extreme outcomes only about 1% of the time in the long run, we would expect to get one or the other first twice as often, or about 2% of the time. A probability calculation in support of rejection of a null hypothesis which takes into account only the likelihood of obtaining by chance an extreme or highly consistent outcome in only one direction (e.g., E superior to C for 13 out of 16 pairs) is called a one-tailed test; one which takes into account the likelihood of getting a similarly extreme outcome in either direction (in the present example, either E better than C 13 times out of 16, or C better than E with the same degree of consistency) is called a two-tailed test. (The terms one-tailed and two-tailed derive from the so-called tails of probability distributions, which are seen graphically in Figure 2 _ to "tail out" in either direction from the central portion of the distribution.)

Considerable argument (mostly, it seems to us, little to the point) has ensued in the statistical literature over the issue of the appropriateness of one-tailed versus two-tailed tests. We shall briefly discuss in a later chapter (Chapter , p. ) what seem to us to be the main points of issue in what might be deprecatingly referred to as "the two-tail cry." For the present we may merely note that either calculation is perfectly legitimate depending on the precise statement of its implications, and that--in any case--the results will differ by only a factor of two in the usual case of symmetrical distributions such as the binomial probability distribution. In saying "only a factor of two" we mean to imply that the main point at issue in interpreting experimental data is generally resolved by theorder of magnitude of the probability, rather than its exact value--e.g., the experimenter will often derive much the same conclusion from a probability value of .02 as from one of .01.

Independence of observations used in a test of significance. One important caution should be mentioned, in concluding this discussion, which applies equally to the data entries used in the sign or binomial test and to those that enter into more sophisticated and somewhat more precise tests such as the $t$ test. This caution relates to the fact that the observations that go into the test must be independent. The concept of independence is discussed further in Chapter 12 as a general problem of sampling related to the basis upon which results may be generalized. Here we only wish to stress the requirement of independence as it relates to the matter internal validity of the test of reliability or "significance."

It will be recalled that a basic assumption of the principles of the probability of compound events (p. 32 ) was that the events were independent. Without inquiring here into all of the factors with respect to which they might be independent or fail to be independent in practice, it is essential to note the basic requirement of independence and to realize some of the important respects in which this requirement might be—and, in fact, often is—violated.

Perhaps the commonest error is to try to use the consistency of a number of different items of behavior observed in the same group of individuals as a basis for a test of significance—i.e., the reliability or dependability of findings. For example, one might use 15 test items in testing the alleged superiority of one experimental treatment, $X_1$, over another treatment, $X_2$. If group 1 were higher on all 15 items one might be tempted to conclude quite erroneously, that the dependability of a hypothesized superiority for $X_1$ had thereby been established. The fatal flaw in this assumption is that the performances of these two groups on the 15 (or more) items are not independent events, since they are the performances, for each experimental treatment, of the same group of people.

To see clearly the fallacy involved here and how seriously it might mislead us, consider the extreme case where the groups are very small—say, 2 pairs of subjects, pair A and pair B, for each treatment group, with one member of each pair randomly assigned to each treatment. In terms of insurance against the chance of happening to get two more able people in one group, the number of independent observations here is clearly only 2, and we would be grossly in error to use a statistical test which assumes 15 independent observations merely because we asked 15 test questions.

The fact that the 15 questions may in one sense be independent items in a random sample from a larger population of items in no way takes care of the crucial issue of sampling error in the assignment of people to treatments: With respect to a basic and obvious source of error in interpreting our findings--namely, that the people assigned to one treatment may be better than those assigned to another--no insurance whatever is provided, per se, by the consistency of results over the 15 items.* Obviously, it would frequently occur that one member (or even both members) of the $X_1$ group would be enough superior to his $X_2$ "opposite number" to do better on all of the items. If, say, one pair ($A_1$ and $A_2$) were about equal, but we merely happened by chance to get, in the other pair, one very bright person, $B_1$, and one quite dull one, $B_2$, we could readily find the performance on all 15 items superior, just as a result of this one chance event. This is not to deny that in the average case the likelihood of getting sizable or consistent chance differences is affected by the reliability of the measures (as determined, among other things, by the number of items in a test). However, the crucial point to realize is that, in this respect as in other connections, the number of independent events that enter into a test of significance must be determined by those features in the conduct of the experiment into which chance is explicitly introduced --in other words, the number of independent acts of randomization performed. In the usual experimental pattern, this is the number of individuals or pairs of individuals (or, more generally, the number of intact groups of individuals or pairs of such groups--see Chapter 10) each of which is independently assigned at random to one treatment or another. In our extreme example this was only two, but the principle applies equally even if the group was sizable. If we had 2, or 6, or 50 randomly assigned pairs tested with our 15 items, the proper respect in which to look for consistency, so far as it concerns the internal validity of an intergroup experimental comparison, is to the consistency of the results for these 2 or 6 or 50 independently assigned pairs--not to the consistency of the 15 test items for one or the other treatment group taken as a whole.*

---

## Chapter 5

# Hypothesis Testing

Introduction

In this chapter the discussion of the nature of experiment is continued with more emphasis on how experiment contributes to the development of "science." In the course of this examination, some of the concepts will be described more formally, or in more general terms, than in the earlier chapters. Thus, the essential concepts to be dealt with have very general applicability and are not limited to experiments on instructional media or any other particular field of experimentation. Accordingly, most of the discussion will, for this chapter, be couched in terms that would apply to any experiment on any subject. The advantage of this is that essential aspects of hypothesis testing by experiment may be seen without restriction to our particular field of investigation. The further application to our specific field of experiments to test hypotheses about the effects of instructional media will be returned to more explicitly later on. For here, let us merely repeat that our purpose is that of contributing to the development of a science of instruction, comprising principles or verified propositions by which some of the guess-work can be removed from the decisions made in designing and using instructional programs.*

Science and the scientific method

The word "science" is used as a short form for "scientific method," or as a name for the process of using the scientific method. "Science" is also used as the label for a body of knowledge of a particular kind.

# Chapter 5

# Hypothesis Testing

## Introduction

In this chapter the discussion of the nature of experiment is continued with more emphasis on how experiment contributes to the development of "science." In the course of this examination, some of the concepts will be described more formally, or in more general terms, than in the earlier chapters. Thus, the essential concepts to be dealt with have very general applicability and are not limited to experiments on instructional media or any other particular field of experimentation. Accordingly, most of the discussion will, for this chapter, be couched in terms that would apply to any experiment on any subject. The advantage of this is that essential aspects of hypothesis testing by experiment may be seen without restriction to our particular field of investigation. The further application to our specific field of experiments to test hypotheses about the effects of instructional media will be returned to more explicitly later on. For here, let us merely repeat that our purpose is that of contributing to the development of a science of instruction, comprising principles or verified propositions by which some of the guess-work can be removed from the decisions made in designing and using instructional programs.*

## Science and the scientific method

The word "science" is used as a short form for "scientific method," or as a name for the process of using the scientific method. "Science" is also used as the label for a body of knowledge of a particular kind.

{Thus, in speaking of a science of instructional media, we mean the development of this (scientific) kind of information about the effects of instructional media.} It is difficult to discuss the word in either of these two senses--method, and knowledge--without reference to the other meaning. The method sense is of prior importance, for it is the method which gives the uniqueness and differentiates scientific endeavor from other creative and intellectual pursuits.

However, the purpose of method is to build a body of knowledge, and it has significance only in the building of a body of knowledge. The scientific method was developed as a means for selecting assertions, sentences, or propositions for the body of knowledge which we will call "scientific." That is to say, there are a large number of propositions considered as candidates for admission to the body of scientific knowledge, but few can be chosen for acceptance. Thus every proposition about the effects of instructional media which can be advanced is a possible candidate, but not all can be selected. Certainly, two propositions which turn out to contradict each other could not be selected. (We say "turn out to contradict" because a real contradiction is not always apparent on the surface in view of the two propositions' having been stated in different language.

To pursue this "candidate" figure further, every man is a possible candidate for political office but not all men can hold office. So there must be some selection method. In our country, candidates are proposed, they must meet some formal qualifications (such as citizenship, age, and residence), and they are placed in competition. The winner of the competition is selected for the office. Similarly, in the scientific method, propositions are proposed and those which meet formal qualifications are placed in competition. The "winners" become part of the body of scientific information. Science, then, is a body of testable propositions which have been selected through the exercise of the scientific method. Propositions to be subjected to the test process are called *hypotheses*. The competition, or test, is set up to choose among *alternate hypotheses*.

### "Truth" and the competition of hypotheses

The democratic process for selecting candidates for political office is actually a fairly good analogy for the scientific method. But why should the scientific method be a "contest"? And is the winner to be decided by vote? Shouldn't it be obvious that the true propositions should be selected, and the untrue rejected? Though this notion seems eminently reasonable, the problem becomes what shall we mean by "true," particularly since what is true often is neither self-evident nor easy to ascertain.

As used in science "true" is a tentative concept. What was considered true yesterday may not be considered true today, and what is considered true today may not be considered true tomorrow. An absolute conception of "true" turns out to be impossible to work with.

If we think of a proposition as being true when it describes or explains a situation "in nature" as it really exists (here, in the "nature" of the effects of instruction on learning), we have a stable definition which should not shift from day to day unless the nature of things itself is changing. Then the way to determine whether a proposition is true or not would be to observe nature (here, instruction and its effects) and see whether the proposition and the observations match. If we could be sure that our methods of observation were infallible, we could make quite direct judgments about the truth or untruth of propositions.

A popular notion holds that the scientific method is a set of techniques for making exact observations. Unfortunately, this is not so. Our observations are subject to many kinds of human fallibility and inadequacies, which (aside from real variability in nature) result in inaccuracies which we call experimental error. (This is true in the so-called "exact" sciences, such as physics, as well as in the as yet quite embryonic science of instructional methods.)

If observation is subject to error, perhaps we should look for other means for knowing truth. Many people have proposed "intuition"--the feeling that a proposition is clearly self-evident--or the opinions of good authorities who have special competencies for judging truth. Such suggestions generally present no real solution because authorities disagree, and one man's intuition does not necessarily agree with a second man's. The dilemma of choosing between competing propositions still exists.

It is, then, difficult to know absolute truth; but maybe we can know the "best" approximation. If we set propositions in competition we may see which is best. The competition is a test to see which proposition will explain or predict the events we can observe in nature more accurately and reasonably. From the competition we may be able to build a science (body of knowledge) which is made up of the most dependable propositions.

Re-evaluation of propositions. Further, just as in the democratic political competition, no candidate is selected to serve forever, and is subject to re-evaluation. Furthermore, the scientific competition may be reopened at anytime, with the same or new candidate hypotheses pitted against each other. Good propositions should be able to withstand repeated tests.

Scientific propositions as statements about relationships. If scientific knowledge is to explain and help us understand nature (including human nature as a special case—and, even more specifically, the effects of instruction on man) then the body of knowledge is going to have to be made up of propositions characterized by more than reliability or even systematic relationships. The eminent philosopher, Morris R. Cohen (1933, p390) puts it as follows:

> The foregoing considerations may thus be put positively: science is a method of testing beliefs on the best available evidence. Logical proof or demonstration, not only in geometry and arithmetic but in all the fields of science, reduces possible doubt as to a large number of propositions to that of a few simple principles. The proof of a proposition makes it impossible to deny it wihtout denying the axioms and other propositions from which it is deduced.
>
> From the nature of evidence it is clear that no one fact can prove another fact unless the two are connected by some constant or invariant relation. Scientific method is thus an effort to make explicit, and to test, the laws according to which phenomena are related to each other to form a system. A dictionary or a railroad time table is not generally regarded as a work of science. For the

> latter must be more than a collection of facts, even if they be organized according to some external order, such as the alphabetic or chronologic one. The ideal of science is to see the facts logically connected according to their essential nature, summarized in some small number of connecting laws or principles.

The scientific status of propositions about instructional media is at present a long way from the ideal as expressed by Cohen. Indeed, no science has fully achieved this ideal. The important point is that the science, at a minimum, is made up of propositions about relationships--not itemization of facts. As a science matures, the propositions approach the ideal more closely. In proposing new "candidate" propositions, we should be working toward the ideal.

Variables. The kinds of propositions (hypotheses) of most interest are more or less sophisticated (in terms of closeness to the ideal) statements of relationships among variables. By a variable we mean here (cf. Chapter 3, p. 10) any characteristic which is used to describe an object or event. This is a simple notion which students often find difficult to comprehend, because they confuse this definition with common colloquial meaning or with more specific technical mathematical usages. When used in the experimental context, "variable" does not carry the connotation of instability or changeableness of common discourse, nor is the term limited to particular kinds of values as required in some specific mathematical formulations. For the present discussion, a variable is any characteristic which is nameable or describable in any way. An apple may be described as red, sweet, soft, ripe and mealy. All these adjectives represent conditions with respect to variables--the variables in these cases being such characteristics as color, taste, hardness, ripeness, and texture (as described in terms of such condition-categories as mealiness-crispness, etc.). They are characteristics by which the apple is described and by which the apple (or class of apples) can be specified or identified.

Similarly, a child's age, his ability to assemble a puzzle, and his height and weight are variables. The length of a film, the number of frames in a self-instructional program, the kind of student activity in a training situation, the number of repetitions of an instructional sequence, the subject matter of a televised lesson, and the order of presentation of a spelling list are also variables. That any characteristic is a variable should not, of course, lead to the supposition that all variables or all ways of describing characteristics are equally useful. Some bases for differentiating among instructional-media variables in terms of usefulness will be subject for discussion in later chapters.

"Qualifications" for hypotheses. Not all propositions, even when stating relationships, qualify for the competition. Like political candidates, hypotheses must meet formal characteristics to enter the contest. In science the requisite formal characteristics are not arbitrary, but are those required for a logical resolution of the test. To qualify, a hypothesis must:

(1) be relevant

(2) be stated so as to imply a test

(3) predict what will happen in the test

(4) be refutable.

The measuring of each of these requirements will be somewhat elaborated presently.

The hypotheses stating relationships among variables take the following form: *If such-and-such conditions occur, then such-and-such results will be observed.* Following the terminology indicated in Chapter 3, those variables which describe the first set of conditions can be called "independent variables." It will also be recalled that variables which are used to describe the resulting conditions are called "dependent variables." In these terms the form of hypotheses can be restated: *If certain conditions with respect to independent variables occur, then certain conditions with respect to dependent variables will be observed.*

In studies of the effects of instructional media, the characteristics of instruction are the independent variables of primary interest, and the dependent variables are the changes in the behavior of the students which it is predicted will result from the instruction. Instructional-media hypotheses will state that for certain conditions of instruction, certain effects, in the form of learning (i.e., changes in behavior), will occur.

### Competition among hypotheses

Although it is common to speak of "*the*" scientific method, there is no single scientific method. Any logical competition among propositions to explain observable events is, potentially, a scientific method. As Wallis and Roberts (1956, p. 5) have put it:

> There is no such thing as *the* scientific method. That is, there are no procedures, formal or informal, which tell a scientist how to start, what to do next, or what conclusions to reach. Scientists rely on the same everyday methods of reasoning that are common to all intelligent problem solving. "The scientific method, as far as it is a method, is nothing more than doing one's damnedest with one's mind, no holds barred." [Bridgeman, 1945].[6] It is

> enlightening, nevertheless, to recognize four stages which recur in intelligent problem-solving, or scientific method.

These four stages are observation, hypothesis, prediction and verification.[8]

From the above characterization of experiments and the discussion of experiment in the preceding two chapters, it may readily be seen that our analogy with the political contest breaks down—not only because no voting is involved, but also because no direct man-against-man competition is involved.

Competitive "scores."[9] The experimental contest among hypotheses is more nearly analogous to sporting events, where each contestant tries to establish a score which represents a measure of his skill, than analogous to direct man-to-man competition. For example, take marksmanship. The firing record must be established under standard conditions, so that a score can be attributed to skill and not explainable by variation of the standards. The difference in scores for two marksmen, both firing under standard conditions, can, presumably, be considered to reflect differences in skill and cannot be attributed to adventitious variants in the standards which enhance one man's score. If one man fired from a position closer to the target than the other man, a higher score might well be accounted for (explained) by the shorter distance rather than superior skill.

In a similar fashion, the experiment gives the hypothesis under test an opportunity to show how well it can predict effects under conditions which rule out alternate hypotheses. The "specified conditions" are introduced for the purpose of ruling out the possibility that the effects which may occur in an experiment could reasonably be explained by hypotheses other than the hypothesis under test. Such conditions are called controls. The hypothesis under examination survives the test and is selected for acceptance when it successfully predicts effects under controlled conditions where the alternate hypotheses cannot reasonably be considered to account for the effects which occur. The analogy between the shooting match and experiment can be made closer if we recast the language we use to describe the contest so that we will be talking about competition among hypotheses rather than among men.

Suppose that from the annual turkey shoot of the country grange, Smith took home four turkeys while Jones took home only two. I explain this difference by the fact that Smith is a more skillful rifleman than Jones. Others contend, however, that a turkey shoot is a carnival where skill doesn't mean much one way or another. Somebody recalls that Jones got a deer last season and Smith didn't. However, for additional reasons (which I may or may not be able to make explicit), I still insist that Smith is more skillful. I, therefore, predict that if placed in a competition firing at a target, Smith will outscore Jones.

Test of competing alternate hypotheses

I will test the hypothesis by arranging such a competition. If, in the competition, Smith makes a better score than Jones, I will consider my hypothesis supported and will conclude that Smith is more skillful than Jones. But I want to be able to say that any difference I may find between Smith's and Jones' scores is attributable to differences in skill, and could not reasonably be explained in any other way. If I had actually had the two men shoot and Smith did get the higher score, what factors other than greater skill could be used to explain the highest score? These other explanations would be alternate hypotheses.

Alternate hypothesis: Smith fired a rifle from a position closer to the target than Jones.

Alternate hypothesis: Smith fired a rifle with a longer barrel than Jones'.

Alternate hypothesis: Smith fired from a prone position while Jones fired from a standing position.

I now have four hypotheses any of which might explain a higher score for Smith. This obviously does not exhaust the possibilities, and I really should try to state as many as I can. (Among these will always be some variant of the hypothesis: Smith made a lucky hit.) Considering at present just these four (the skill hypothesis plus the three stated alternates), how am I to choose among them? I set up the test situation to get evidence which will permit me to make an inference about the skill hypothesis, which is the one I choose to test, and control against the alternate hypotheses. If both men fire from the same distance from the target, the rifles are of the same type, and both men fire from the prone position, I have ruled out at least the alternate hypotheses stipulated.

It might perhaps now be conceded (still for the present ignoring other quite possibly important alternate hypotheses) that I now have arranged conditions so I can select the skill hypothesis as the best explanation if the prediction from this hypothesis does occur.

This is the nature of the competition among hypotheses and is the nature of an experiment which permits a choice among alternate hypotheses.

Formal properties of hypotheses

Earlier in this chapter, we listed, without definitions, four characteristics which hypotheses must have to qualify for the competition among hypotheses. Let us now try to define these properties and examine why they are required for a logical resolution of the task of hypothesis testing.

(1) A hypothesis must be relevant.

When we are concerned with some problem or question, we are not interested in propositions which are not answers to the question. The answer may be good or bad--that is what we want to determine by our experiment--but there must be some reasonable possibility that it can be an answer. Irrelevant hypotheses--those which cannot reasonably be proposed as answers to our inquiry--are not in competition. The winner of the 440-yard race is not selected from among the field of contestants in the mile race. Hypotheses about methods for effective instruction on the assembly of a carburetor are not relevant to explaining why Smith shot a better record than Jones, and they are not entered in that competition. Hypotheses about what form of film will be most entertaining on television are not, *per se*, relevant to the question of what form of film will best teach facts in a classroom.

(2) A hypothesis must be stated so as to imply a test.

This is to say that there must be some way that an appropriate test situation can be logically derived from the hypothesis, or we could not know what kind of competition to put it in. When we say that Smith is a more skillful rifleman than Jones, the test implied involves firing rifles and not a handball game. Another way of justifying this requirement is that if the test situation could not be logically derived from the hypothesis, there could be no means of logically relating the outcome of the experiment to a statement about the goodness or badness of the hypothesis.

In non-trivial situations, the test is not as directly obvious as in the case of Smith vs. Jones. Consider the proposition that material presented aurally will be remembered longer than material presented visually because the aural material reverberates in the brain while the visual material does not. This may be thought to be a wild proposition to which few people would give serious consideration. If you did want to test this proposition, how would you observe the reverberations? There is no way we know of, but that does not mean the hypothesis cannot be tested. The hypothesis could be tested if we could derive some logical consequents of the existence of these reverberations which could be observed.

The following example is given by Cohen and Nagel (1934, p. 208):

> Thus the hypothesis that the sun and the planet Mars attract each other proportional to their masses, but inversely as the square of their distances, cannot be directly confirmed by observation. But one set of consequences from this hypothesis, that the orbit of Mars is an ellipse with the sun at the focus, and that therefore, given certain initial conditions, Mars should be observable at different points of the ellipse on stated occasions, is capable of being verified.

That the test situation may not be directly obvious is one reason, among others, that defining experiments involoves considerable ingenuity.

(3) A hypothesis may explain all known facts very well but, to be of serious interest, it should predict new data which has not yet been observed--it should predict what will happen in a test consisting of an experiment which has not yet been performed.

Of course we will be happier with hypotheses which lead to a new knowledge because that is, after all, the business we are in; but there are more elementary reasons for requiring prediction. First, if we explain data after the fact, we may formulate an hypothesis which just happens to explain the data we just happened to have. In the Smith-Jones argument, the hypothesis that Smith has more skill can explain his getting more turkeys. On the other hand, the opponents argue that the turkey shoot was just one particular set of data that I chose to talk about--what about the results in the deer season? They might argue, further, that there are many other hypotheses, not involving skill, which could also explain Smith's getting four turkeys. Then we still have the problem of choosing among theses hypotheses. It is often the case that two or more equally good hypotheses can be formulated ex post facto. A much more interesting reason for asking for prediction is that some of the most important experiments in the history of science have been "tie-breakers" for deciding between two hitherto equally plausible hypotheses.

(4) To be testable, a hypothesis must be refutable. We want a hypothesis to explain what happens and to explain as much as possible of what may be observed. However, if a hypothesis can explain anything which could possibly happen, there is no sense in trying to test it in competition with other hypotheses, because it can't lose.

"When he is 'on,' Smith can out-shoot anybody" is an unacceptable hypothesis--i.e., untestable--because there is no way of refuting it. Whatever may occur in the test, the hypothesis can still be asserted. If Jones makes the higher score, we just say Smith wasn't "on" today.

Consider what would be an appropriate test for the assertion that, "If the subject matter is presented clearly enough, any student can achieve mastery." Are there any conceivable conditions which would make it possible to say that some other hypothesis is better--that this one is untrue? Even if no student in the test situation can achieve mastery, we can say the hypothesis is still true--the subject matter just was not presented clearly enough. A towering champion may be expected to win every race, but he can't be considered the champion unless there is open competition with the possibility that somebody else can win. A good hypothesis may continue to predict the relevant range of facts better than any competing hypotheses; it is essentially untestable unless it could potentially be overthrown or refuted (or at least fail to be supported) by the evidence.

Bias and control

Setting standard conditions for the marksmanship competition, or other contests, is usually justified on the basis of "fairness." It is not fair for one contestant to fire under more favorable conditions than another. It is not fair for one poker player to have six cards while the others have five. Strictly speaking, we are not interested, in experiments, in setting conditions which will be fair--but in avoiding bias. The notion of "fairness" does serve rather well as a descriptive analog for "avoiding bias." However, the experimenter is not really interested in sportsmanlike treatment of hypotheses, but rather in conducting his work so that he may make logical statements about what he did and what he observed.

It should be remembered that every experiment to test a hypothesis about effect will have at least two treatments. If some independent variable, other than the treatment variables, systematically operates more heavily in one group or treatment than in another, then we have bias. In the presence of bias, we cannot know whether the effects should be attributed to the treatment variables or to other variables unless there is some way of assessing the extent of the bias.

We may differentiate two kinds of contaminating or extraneous influences: systematic influence, or bias, and random influences--also called systematic error and random error. Systematic bias is error which operates in a consistent (though possibly unknown) manner in our experiment but differently for one treatment than for others. Random influences operate in a chance fashion--sometimes favoring one treatment, sometimes another.

"Controls" function to eliminate the effects of systematic bias. Random influences cannot be eliminated, but valid inferences can be made about the extent of their probable effect in an experiment which includes randomization.

An "unbiassed" or unprejudiced comparison is merely one which lacks systematic bias; evidently no comparison can be immune to random influences. Validity in an experiment is attained when systematic bias is removed (either by making it non-systematic through randomization, or by procedural control so as to satisfy the experimenter and his peers that any inequality is inconsequential) and when random influences are such that the extent of their influence can be validly assessed.

### Experimental design and experimental pattern or structure

We digress here to discuss the term "experimental design." We may use it with the specific meaning of the strategy for controlling alternate hypotheses. Through the literature of experimental inquiry, however, the term has a variety of meanings. Sometimes a research report will have a section captioned "Experimental Design." Often, such a section will describe the experimental design (in the broad sense in which we will use the term), including all the information about the conduct of the experiment that the author intends to give--its procedures, structure, measures, numbers of subjects, etc. In other instances, especially in statistical (or statistically oriented) textbooks, "experimental design" is used more narrowly to mean only the pattern or structure of the experiment which determines the procedure for estimating experimental error. In general we prefer to follow Tukey (1960, p. 81) in differentiating research design from research pattern or structure--using "design" in the broader sense and "pattern" or "structure" for that aspect of the experiment concerned with statistical methods for estimating experimental error:

> "Design of experiments" has been used too long in statistics to describe a process which is only a small part of the whole process of designing an experiment. From time to time, this conference has gone so far as to use the term probability arrangements for patterns of experimentation in which randomization plays a major role in simplifying the interpretation of the observations (such patterns as randomized blocks and Latin squares) and the term

informative patterns for patterns of experimentation whose main aim is to get us the information we want, and might otherwise miss (such as surface fitting designs of first or second order........ This use of terms was a forward step. We must support and continue it. For only as we recognize the true, limited scope of the contribution of randomized and informative patterns to design, only as we become willing to recognize ourselves as "patterneers" rather than "designers" while we are carrying out only this limited function, can we hope to sell these techniques, which are indeed useful, to those scientists and engineers whom we would most like to see adopt them.

We did not, I am afraid, face up to the whole of this change, the necessary freeing of "design" for its broader meaning. Yet we must. For the design process is like a long staircase, where "probability arrangements and informative patterns" stand up on the bottom step. They are there because they are chosen last, after the subject of experiment, the variables concerned, their levels or versions (number, nature, and, if quantitative, mode of expression), and all the other matters of "importance" have been chosen. While "probability arrangements or informative patterns" may make or break an experiment, they are minutiae from the broad view of designing the experiment itself. They never ought to have come to be called "designs," if only because of the danger that their student may think he can "design" an experiment when he can only "pattern" it.

## The role of statistics in experiments

A reason for this differentiation between "design" (the more inclusive term) and "pattern" or "structure" (the more restrictive terms) is to emphasize that statisticizing and experimenting are not co-extensive. Contrary to the impression one may get from some texts with the term "experimental design" in the title, actual computation of statistics has but a relatively small, albeit important, role in the actual conduct of experiments. The experimental process includes everything from formulating the original question to stating a conclusion, and should especially emphasize control. Though the experimental design and the statistical pattern must match (cf. Chapter 3 , pp. 35 ), we must not start with a statistical pattern and select controls to go along. The design of controls is prior to the selection of a statistical pattern. Nevertheless, the design must result in a pattern for which techniques for estimating experimental error exist or talent is available for deriving a new method. The concordance of design and pattern must be established before collecting data.

The role of statistics and its relation to experimental inquiry has been summarized as follows by Wallis and Roberts (1956, p. 5)

> Statistics provides rational principles and techniques that tell when and how judgments can be made on the basis of this partial information, and what partial information is most worth seeking. In short, statistics has come to be regarded, as we said in the first sentence, as a method of making wise decisions in the face of uncertainty.....
>
> Statistics is not a body of substantive knowledge, but a

> body of methods for obtaining knowledge. As such it should be viewed against the background of general methods of obtaining knowledge--of general scientific method, in short.

Statistical consultants

The authors would recommend that suitable statistical consultants be used for this purpose when at all possible, for it is a large order, which very few people can meet, to be an expert in your field (instructional media) and in statistics too. Lubin (1962) screened 1400 articles for his review of statistics in the Annual Review of Psychology to cover a two year period. Michael and Hanka (1960), covering a three year period for their review of statistical methods in Review of Educational Research, say, "More than 1200 references were located, and about 450 have been included..." It is evident that the neophyte investigator, or even the average experienced experimenter, can hardly expect to keep up to date with this literature. It should be noted, however, that the statistician can be expected to be an expert in statistics and not necessarily in experimental design. He is even less likely to be an expert in your field, and productive experiments are not to be expected without this special-field expertness.*

It is for this same reason that this book devotes little space to the elaboration of statistical techniques per se. The basic logic of statistics, however, must be in the repertoire of every experimenter. The relationship between statistics and experimental design (in the narrower sense of experimental structure) has been most cogently expressed by Fisher (1960, Sec. 1, pp. 1-3) in discussing the grounds on which evidence is disputed:

When any scientific conclusion is supposed to be proved on experimental evidence, critics who still refuse to accept the conclusion are accustomed to take one of two lines of attack. They may claim that the *interpretation* of the experiment is faulty, that the results reported are not in fact those which should have been expected had the conclusion drawn been justified, or that they might equally well have arisen had the conclusion drawn been false. Such criticisms of interpretation are usually treated as falling within the domain of *statistics*. They are often made by professed statisticians against the work of others whom they regard as ignorant of or incompetent in statistical technique; and, since the interpretation of any considerable body of data is likely to involve computations, it is natural enough that questions involving the logical implications of the results of the arithmetical processes employed should be relegated to the statistician. At least I make no complaint of this convention. The statistician cannot evade the responsibility for understanding the processes he applies or recommends. My immediate point is that the questions involved can be dissociated from all that is strictly technical in the statistician's craft, and, *when so detached*, are questions only of the right use of human reasoning powers, with which all intelligent people, who hope to be intelligible, are equally concerned, and on which the statistician, as such, speaks with no special authority. The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation.

The other type of criticism to which experimental results are exposed is that the experiment itself was ill designed, or, of course, badly executed. If we suppose that the experimenter did what he intended to do, both of these points come down to the question of the *design*, or the logical structure of the experiment. This type of criticism is usually made by what I might call a heavy-weight authority. Prolonged experience, or at least the possession of a scientific reputation, is almost a prerequisite for developing successfully this line of attack. Technical details are seldom in evidence. The authoritative assertions "His controls are totally inadequate" must have temporarily discredited many a promising line of work; and such an authoritarian method of judgment must surely continue, human nature being what it is, so long as theoretical notions of the principles of experimental design are lacking--notions just as clear and explicit as we are accustomed to apply to technical details.

Now the essential point is that the two sorts of criticism I have mentioned are aimed only at different aspects of the same whole, although they are usually delivered by different sorts of people and in very different language. If the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too. It is true that there are a great many experimental procedures which are well designed in that they may lead to decisive conclusions, but on other occasions may fail to do so; in such cases, if decisive conclusions are in fact drawn when they are unjustified, we may say that the fault is wholly in the interpretation, not in the design. But the fault of interpretation, even in these cases, lies in overlooking the characteristic features of the design which lead to the result being sometimes inconclusive, or conclusive on some questions but not on all. To understand correctly the one aspect of the problem is to understand the other. Statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation.

## Control techniques

Sometimes it is said that the essence of experiment is manipulation. This is a valid statement in the sense that the experimenter determines and creates the conditions of the test. He makes the test situation, and the success of the experiment depends on how well he does this. But sometimes "manipulation" is also used as synonymous with control. To the present authors, it would seem more useful to differentiate these two terms and reserve the use of the word "control" to the specific meaning of ruling out alternate hypotheses. Control may require manipulation, but there is much more manipulation by the experimenter. He assigns treatments, he introduces variables, and uses particular values of variables.*

The confusion of "control" and "manipulation" may be due to the fact that achieving control involves manipulation of independent variables, or having an operational equivalent of manipulating the independent variables. All the alternate hypotheses propose that some independent variable will have some effect. The way to control a hypothesis is to keep such extraneous independent variables from influencing the experimental results. Several general techniques or schemes for doing this can be described.

Only one of the control techniques, randomization, occurs in every experiment. Randomization will automatically control for many, though not all, alternate hypotheses--but it is not necessarily always the best method.** Except for the fact that randomization must be used in every experiment, the design of an experiment is the formulation of schemes for controlling alternate hypotheses. This is a creative act and no standard procedures can be specified in a "cook book"--or, if you please, handbook. For some particular kinds of independent variables it can be said that a particular kind of control is the standard technique in the field, or even that it is the only good technique that anybody has ever come up with. But this is generally not the case. Usually, more than one way will work, and the experimenter must make a choice in designing his experiment. It should also be said that one would not necessarily always choose the most effective control techniques. For instance, every time randomization is used to control an independent variable, the experimental error is larger that it would have been if another available control technique were used. Nevertheless, the experimenter will often elect to control through randomization.

Experimental hypothesis testing

We are concerned with the experimental verification of some hypothesis we have chosen to test. To be amenable to test by experimental procedures, the hypothesis must predict that there will be some differential outcome (values of the dependent variable or variables) which will ensue from two or more identifiably different conditions or combinations. (As previously noted, the several combinations of independent variables are often called treatments.)

If the hypothesis does not directly state such a prediction, it can be tested through a prediction which can be shown to be a logical consequent of the hypothesis. The hypothesis, either the original or the logical consequent, which is stated in the form which specifies the experimental conditions and the expected effects, is called the *experimental hypothesis*. The independent variables as defined in the experimental hypothesis can be termed the *experimental variables* (cf. also Chapter 3, p.          , and Chapter 7, p.          ).

The alternate hypotheses are also statements of relationships which state that the same kind of effect which is predicted in the experimental hypothesis can be expected to occur as effects from other independent variables. To be able to attribute effects to our experimental variables and not to extraneous variables (the independent variables of the alternate hypotheses) we must control extraneous variables;

The effects which may ensue from (or be caused by) the independent variables of the alternate hypotheses will obfuscate the logical processes of attributing observed effects to the experimental variables only if these effects are (1) systematic, (2) differential, and (3) inseparable from the experimental effects. If we can destroy any of these three properties we will have achieved control.

The systematic property can be destroyed through randomization. Possible biassing influences are not prevented, but they will appear as random error rather than systematic bias. Random error can be dealt with after the collection of data, but we had best deal with systematic bias before the collection of data.

The differential property can be avoided by assuring that the same values (amount or quantity) of each of the alternate independent variables act upon each of our experimental treatments. We can use the term *exclusion* when the value is zero, and the term *equating* when the values are other than zero (though possibly not known). *Stabilization* refers to a technique which requires some operations other than, or in addition to, selection. Counterbalancing can also be used when treatments or the measurement of treatment effects are sequential.

Note that it is possible to know that two values are equal without knowing the size of either value. Note also the difference in treatment of the experimental and the alternate independent variables. We could say that we try to prevent the alternate variables from introducing bias while we are trying to give the experimental variables every opportunity to introduce bias.

If one of the above controls has been used, the effects which might be due to the alternate independent variables may or may not have been rendered separable. We need not necessarily take further action against this property of inseparability. However, we may want to do so to improve the sensitiveness of our experiment.

If we have no other means of control we must use some technique which will yield an estimate of the amount of the effect which should be attributed to the alternate independent variables. The common technique involves the introduction of a control group or groups. When using a control group, we do not exclude the bias due to alternate or control variables from any of the groups. Rather, we can get measures of effects due to a combination of the experimental and the control variables for one group and of effects due to the controlled variables alone for another group. Subtraction, in effect, of the second from the first can then give us the answer we want--namely, the effect of the experimental variables.

The term "control" is used in so many different ways, and often in such unfortunate ones, that a great deal of confusion revolves about the word. We have tried to limit the term to its technical and most useful meaning. When coupled with "group" to form "control group" the confusion is compounded. Many people say that there is no experiment without a control group. Unless they also say or imply that every group in an experiment is in a (quite defensible) sense controlling against every other group, the statement is incorrect. Usually, they mean that an experiment must have a group from whose treatment the experimental variables excluded may be used by choice, but it is required only when it is the only means for controlling against some alternate hypothesis. In our discussion we limit the usage of the term to groups which are introdcued into the experiment for control purposes.

*When one is reading research reports it is essential to determine how the term "control group" is used in that particular paper.* Sometimes, as in cross-media comparisons, the control group is that which uses the "conventional instruction" as differentiated from those students who are taught by the proposed new techniques. In other instances, it is a group which received *no* instruction, and thus furnishes a baseline from which the effects of instruction may be measured. In still other instances, it represents a particular form of instruction from which certain specified features of experimental treatment are excluded.

## Concerning the null hypothesis

It might be said that after we have done all we can to rule out alternative hypotheses by controlling against them, we still have one alternative hypothesis left because we can't control against it by any procedural means within an experiment. This is the hypothesis that the effect we found can be accounted for by chance. This is the hypothesis is the null hypothesis. (cf. Chapter 3, p. 25. )

Since there is experimental error, and it can never be wholly removed, we must expect that there will be observation which look like real effects even though our independent variables in truth have no effect. That is, if we could free the observed effect from the effects of chance (experimental error) we would see that the effect of a treatment might truly be zero.

Now, we cannot say how much chance is operating. But if we assume explicitly that some chance factor, not our experimental variables, is operating to yield effects which differ from zero, we can in a properly designed experiment estimate the likelihood of getting effects as large as (or larger than) we found. If the likelihood is very small, we may be willing to reject the null hypothesis, thus deciding in favor of the hypothesis the experiment was designed to test.

The critical point here is well stated by Wallis and Roberts, (1956, p. 8), who also make the additional important point that the hypothesis under test may itself state only a probabilistic rather then an absolute or unvarying prediction:

> In checking a prediction with new numerical data, it is crucial to realize that the data and the prediction can seldom be expected to agree exactly, even if the theory is correct. Discrepancies may arise simply because of chance circumstances ("experimental error") that are not inconsistent with the theory. Furthermore, many important theories of modern science are probabilistic or "stochastic" rather than deterministic, in that they do not predict precisely how each observation will turn out, but only what propotion of the observations will in the long run turn out in each of a number of possible ways. Genetic theories, for example, do not in general specify the

> characteristics of each individual offspring of a given parentage, but only the proportions in which certain kinds of offspring will appear. Such theories, furthermore, do not specify the proportions for any one set of observations, but only the "long-run" proportions or probabilities. In comparing a set of observations with theory, the question to be considered is, therefore, "Is the discrepancy reasonably attributable to chance?" If the discrepancy can reasonably be attributed to chance, the theory is not contradicted, and there is no adequate reason to seek special "causes" to explain the discrepancy. If the discrepancy cannot reasonably be attributed to chance, it is appropriate to look for causes--that is, to modify the theory.
>
> Modern statistical reasoning has given a definite meaning to the verification of a hypothesis. A hypothesis is verified--"tested" is perhaps a better word--to the extent that the influence of chance in the evidence has been correctly interpreted. Statistical procedures have been evolved for measuring the risk of incorrect interpretation objectively, in terms of numerical probabilities; or, to put it differently, for measuring the risks of erroneous conditions.

Notice that we do not rule out, or control against, the hypothesis by preventing chance to operate or forcing chance to affect all our treatments equally. In rejecting the null hypothesis we say only that we are going to talk about it as being false. We could be wrong. It could be true. However, if we have a proper estimate of experimental error we can specify the risk of being wrong. The level of risk we may be willing to take is called the "Level of Rejection."

The level of rejection, or confidence level, is chosen by the experimenter. When he adopts a 1% level of rejection, he is saying that he is willing to take the risk of being wrong in this kind of situation one time in a hundred. A 5% level of rejection involves a risk of rejecting a true null hypotheses 5 times in a hundred.*

How is the level of rejection determined? Unfortunately, there is really no very good answer. One kind of answer is "by convention"; the level of rejection that researchers in this field usually use. Insofar as the convention reflects the accumulated experience on useful levels, this is not a bad reason. Insofar as it reflects "inflation," the conventional level should be considered too risky. When the authors of this book were young, the conventional level of rejection was .03%. As inflation set in, the level became 1%, and nowadays 5% is becoming more and more common.

Though subject to considerable vagueness, there is a more logical way to select a level of rejection. We accept a more stringent or more lenient level of rejection depending on how well asserting the effects as resulting from the experimental variables fits with other known information. If the proposition is quite compatible with other currently accepted propositions we may be willing to take larger risks than for a proposition which if asserted would call into question a large quantity of currently accepted knowledge. (See also p. 36.)

Some people argue that this argument brings in prior knowledge, and an experiment must be a self-contained event. Here a statement by E.B. Wilson (19 , p. 59-60) is highly pertinent. Wilson says:

Hardly any experiments are carried out in a complete vacuum, and so it is seldom that no prior knowledge can be brought to bear on a given problem. There has been much dispute as to how this knowledge should be utilized... It is rare that it can be given quantitative expression. Nevertheless it is folly to ignore it completely, and this is in practice not done. One way of bringing it into the picture is to use it to help set the permissible risks of the two kinds of errors. Thus if all prior knowledge strongly supports the null hypothesis, the risk of wrongly rejecting it should be made small, and vice versa.

It will undoubtedly be objected that since quantitative expression cannot be given to prior knowledge, it should not be used to answer a quantitative question, such as the choice of a level of significance. However, a choice of significance level does have to be made if any decision is to be reached as a result of the experiment. This will inevitably be a somewhat arbitrary exercise of judgment in the best of circumstances, and it seems reasonable to include in the bases of this judgment all the pertinent information.

Books on statistics are likely to dismiss this question with the statement that usually 5 per cent or sometimes 1 per cent is taken as the level. A more realistic statement would seem to be that a level somewhere between 40 per cent and 0.0001 per cent will be appropriate for most cases, the exact value being the result of considerable thought on the part of the investigator concerning all the items discussed above. Furthermore, even though in some cases definite yes or no decisions have at some point to be taken even though the results are just barely above or below the level of significance

> decided upon, ordinarily it is wise to set upper and lower limits of significance with a no-decision zone in between. Results falling in this zone would indicate the need for further investigation.
>
> Unfortunately, a hundred repetitions of a biased experiment is merely one hundred times more misleading than one biased experiment. In fact, in cases where results are obtained which seem highly unreasonable on the basis of prior knowledge, instead of requiring repetitions until the level of significance is one in a million or so, it is usually more frutiful to comb the experimental design and procedure for possible sources of bias before continuing.
>
> It should be emphasized that the risks to be run should be set in advance and not after the results are in. It is too easy to shrug off an unpopular result, after it is obtained, by readjusting the risk levels. It is difficult to get people to agree to act in definite ways depending on the results of a future experiment, but if no other evidence is to become available and no criticism is made of the experiment, there seems no valid reason for not making all the intuitive and judgment decisions beforehand and introducing them by way of the risk levels.

When the risk involved in rejecting the null hypothesis is too great, we do not reject it. That's all we do; we do not reject it. We do not say that it is true; we just don't reject it. In not rejecting, we are saying that the effects can be accounted for by chance factors, but we have done nothing to control against the hypothesis that the effects are due to our experimental manipulations. That hypothesis is still in competition. We cannot choose between the two hypotheses on the basis of the experiment.

"Significance." The evaluation of the null hypothesis has unfortunately come to be known as "significance testing." A difference is said to be significant if the null hypothesis is rejected. This leads to all kinds of mischief through confusion in the use of "significant." Sometimes, "significant" is interpreted as meaning important and, sometimes, sentences are formed using the word "significant" as though it applied to the experimental hypothesis, or both.

"Significant" here does not mean anything like "important." In fact, the real meaning is nothing like any common usage of the word significant, and talking about "significance" does not add any information to a statement of rejection or non-rejection of the null hypothesis. It would be best if we could just forget the word. If an adjective must be used to describe qualitatively the outcome of an experiment in terms of the rejection level which is attained (or which fails to be attained), the term "reliable" or "dependable" (or "unreliable" or "undependable") seems preferable on grounds of being less misleading. Generally, however, we feel it is better to state the finding operationally in terms of the level of probability that the observed effects could be ascribed to chance factors.

When the "significance" statement is slurred over to apply to the experimental hypothesis, it is often inferred that the difference, if any, is unimportant, or even that there is no difference. There either were or were not greater effects for one treatment than another. When we cannot reject the null hypothesis, we do not know how to explain any differences which may have been found.

Experimental hypothesis and null hypothesis. Often the reason that the significance statement is applied to the experimental hypothesis is the use of the erroneous practice of stating the experimental hypothesis in the form of a null hypothesis. They are and should be separate hypotheses. That is we need to distinguish between the status of an experimental hypothesis and that of the null hypothesis, which implies no directional prediction. And we would argue against the stating of experimental questions in terms of a null hypothesis. The null hypothesis is something that comes later as a means whereby an experimental hypothesis may be accepted through refutation of the null hypothesis. It is not uncommon to read research reports which start with the author's explanation of why he thinks that treatment A will have an effect B and then in the next paragraph he states as his "experimental hypothesis" that treatment A will have no effect. No! The experimental hypothesis says that there will be an effect (and may predict its direction). The null hypothesis says that any observed differences are experimental error. The experiment is performed to choose, if possible, between these two hypotheses. The choice is always one-sided however; we can reject the null hypothesis, but if we fail to do so we cannot reject the experimental hypothesis. However, when the experimental hypothesis predicts effects in a specified direction, it can be rejected if the results come out in the opposite direction to such a degree that the null hypothesis is also rejected.

Finally, dismissing any notion that the point can be overly labored, we stress again that the central, basic idea which must govern the thinking about experiment is that the purpose of scientific method is to choose, if possible, among alternate hypotheses. Many errors in research will be avoided if this principle is kept in mind. It should not be implied that if you don't reject the null hypothesis, you accept it!

### Decisions versus conclusions

The rationale of confidence levels in hypothesis-testing experiments, as related to the kind of question an experiment is addressed to, may be illuminated through some distinctions made by Tukey (1961).

Tukey makes a distinction between decisions and conclusions. Decisions are essentially involved in practical matters where a course of action made must be decided on, wheras conclusions are matters of science or knowledge which can be accepted as having been proved or established beyond a reasonable doubt. The history of science necessarily has engendered an emphasis on skepticism, on suspending judgment in absence of reasonable certainty (there is never complete certainty), and hence on demanding rather rigorous proof for accepting conclusions and perpetuating them as demonstrated or demonstrable truths in the literature of the science.

In directly and immediately practical matters, by contrast, the characteristic requirement is to be able to decide upon a course of action, often in the face of grossly incomplete evidence--indeed, in many instances where the decision must be based on evidence so incomplete that there is barely a fifty-fifty chance that the decider will be right.

### Practical decisions versus scientific conclusions

There are situations in respect to military decisions, business decisions, and decisions of personal life, where a "batting average" of .500 is not only a good record in making correct decisions but almost an unprecedentedly good one--as is of course true also in baseball, where even in the major leagues a batting average of .400 is phenomenal. On the other hand, when we want to arrive at conclusions which will be perpetuated as a basis for many recurrent decisions in the future, intuition as well as the history of science may lead us to the conclusion that we must insist on a much higher probability of being correct than merely having a relatively good "batting average" or even of being right considerably more often than wrong.

Thus we are led to insistence on a moderately or very stringent degree of confidence as exemplified in the setting of rejection levels at the .01 or .05 level for the rejection of a null hypothesis in an experiment. However, it needs to be noted that in the course of deciding what to experiment on or how to conduct an experiment we are also frequently faced with decisions where the basis for deciding what course of action to adopt--e.g., in selecting or conducting or analyzing the data from an experiment--must be based on much less confidence.

Decisions

Tukey distinguishes, in the case of decisions, three alternatives in situations where a decision must be made between two possile alternative courses of action, where one "must choose between two policies...[for a situation] that is now upon him," in which case he "must weigh alternative A against alternative B in this immediate situation, and strive to select the alternative that will yield the bigger reward, whether this reward be a cheaper safe bridge, a better chance of recovery for the patient, or a more profitable season." Tukey schematizes this in the following manner: "In each instance the individual must judge whether to act as if the reward from alternative A will indeed prove to be greater than that from alternative B, (which we may abbreviate 'A greater than B'), or whether the opposite is true ('A less than B')."

In this case the three alternative decisions can be stated as follows:

(1) To act in the present situation as if A is greater than B.

(2) To act in the present situation as if A equals B.

(3) To act in the present situation as if A is less than B.

Tukey stresses the importance of realizing that when we say "act as if A is greater than B", we have made "no judgment as to the 'truth' or 'certainty beyond a reasonable doubt' of the statement 'A is greater than B'. When we say 'for the present', we are referring only to the particular situation under consideration at the present."

Conclusions

The situation in the case of accepting a conclusion which we proposed to add to the body of "scientific knowledge," on which many future decisions may be based, presents quite different requirements. Here a conclusion such as A is greater than B is to be accepted and it is, as Tukey points out, "taken into the body of knowledge not just into the guidebook of advice for immediate action as would be the case with the decision, it is something of lasting value extracted from the data."

This is true even though the conclusion is always accepted with some tentativeness and subject to future rejection when and if the evidence against it should subsequently become strong enough. But since we are to use the conclusion as the basis for a host of subsequent decisions we are much more cautious in accepting it than we are in making a single immediate practical decision. For should it be controverted, not merely a single decision but potentially a whole series of decisions that may have led to complex and costly actions are then shown to have been based on faulty premises. The acceptance of the conclusion as "of lasting value" may seem paradoxical in the light of our previous comments (p. 3 ) about the tentativeness of scientific conclusions. Perhaps this paradox in the nature of the acceptance of a conclusion in science can be summarized by Tukey's statement that "it is taken to be of lasting value, but not necessarily of everlasting value."

Relation to purposes of an experiment

The foregoing needs to be elaborated somewhat in terms of the distinction between practically oriented or technological experiments for product-testing, which involve immediate decisions to accept or reject a product, or to accept or reject a particular element (e.g., a frame or series of frames in a self-instructional program). These can be contrasted to experiments designed to test hypotheses or verify principles that will apply to a large number of subsequent actions or decisions. The implications are that for experiments designed to test hypotheses we need to insist on much more rigorous standards. The increased rigor applies to both of two considerations--first it applies to the selection of a rejection level; here we may, for many purposes, well wish to suspend judgment unless a hypothesis is verified by rejection of the competing null hypothesis at a confidence level in the .01 to the .05 region. Second, it applies also to the rigor of control in the experiment. For a practical decision concerning the revision of a particular instructional sequence, we may decide to dispense with certain formal controls, relying on informal observation and intuitive judgment. This is, indeed, a practical necessity, since we obviously cannot conduct a carefully controlled, rigorous experiment to make each little decision that enters into the building of an instructional sequence (or into the writing of a book such as this one).

The difficulty in resolving the difference in requirements here as to how good evidence we wish to insist on lies in the fact that most of our decisions lie in between "gray" regions, not as clear-cut as the testing of an important theory of very general applicability, at the one extreme, nor of some small immediate practical decision, such as how to construct a sentence or write a single frame in a self-instructional program, at the other extreme.

Obviously for the latter situation we would not do a formal experiment at all. Perhaps the greatest uncertainty will arise within the range of formal experiments—that is, experiments set up in the general model of a controlled experiment, but where neither the research sponsor, the experimenter, nor the reader of the results may have a clear basis for deciding the status of the conclusion that is to be arrived at—that is, whether the outcome of the experiment (A greater than B) is to be regarded as the basis for a scientific conclusion "of lasting value," or whether it is only a basis for a single practical decision, or whether it is somewhere in between—e.g., a basis for deciding whether to make a series of films in color or in black-and-white, or whether to contract for the construction of a series of self-instructional programs in "Crowder" form or in "Skinner" form, or whether to expend a major share of a school's budget in the installation of a closed-circuit television versus putting more money into trying to recruit and train more and better teachers.

Simple versus complex experimental hypotheses

It should be noted that hypotheses which we may wish to test can differ considerably in complexity.

Unqualified hypotheses. The simplest form of experimental hypothesis is an "unqualified" one. It simply asserts that some variable $X$ (which will be manipulated experimentally to test the hypothesis) will have an effect on some dependent variable, $Y$--for example, that showing a given film will increase students' interest in biology, or that a given teaching-machine program will teach students how to factor quadratic equations (or at least measurably improve their ability to do so).* The prediction is, in effect, that $Y$ will be greater than zero (i.e., in symbols, $Y > 0$ ).

The basic logic or form of such an unqualified hypothesis is essentially the same if, instead of merely predicting an effect of $X$, we instead predict that some category of variable $X$ (say, $X_1$) will be superior to some other value of category of $X$ (say, $X_2$). For example, we may hypothesize with respect to color (variable $X$), that a color film ($X_1$) will be better than a corresponding monochrome print ($X_2$). (Cf. Chapter 3, p. 274.) The prediction here is that amount learned ($Y_1$) from the color film ($X_1$) will be greater than the amount ($Y_2$) learned from the black-and-white film ($X_2$); in other words, that $Y_1 > Y_2$. However, this is equivalent to saying that the difference $Y_1 - Y_2$ is greater than zero ($Y_1 - Y_2 > 0$), so that, by merely considering $Y_1 - Y_2$ as the dependent variable, the form of the hypothesis predicting the effect remains the same as before. It may also be observed (as further discussed in relation to "control" in experiments) that, in any case, showing that $X_1$ has an effect that will always require some comparison conditions (not-X, or $X_0$ --or call it $X_2$, for that matter); that is, in any case this simplest prediction, merely that some effect was produced, will require showing, in effect, that $Y_1 > Y_2$ or $Y_1 - Y_2 > 0$.

Qualified (contingent) hypotheses. For various reasons (to be discussed later), we are more often interested, in instructional-media experiments, in contingent or qualified hypotheses than in unqualified hypotheses. A qualified or contingent hypothesis states a prediction that X will affect y in a given direction if condition A obtains, or, more likely, that the effect of X on y will be greater if A than if B. Here, A and B are qualifying independent variables. For example, we may predict that the superiority attained by use of color in a film (or, some form of "branching" in a teaching machine program, etc. ), as against some comparison treatment (monochrome film, "linear" program, etc.), will be greater for boys than for girls, or that the amount of advantage will be greater if students are bored or disinterested (i.e., will make more difference under these conditions than if the students are interested or alert). In this case the qualifying variables are pre-existing (pre-experiment) variables which may be classed as "population" or learner characteristics (if taken as found by sorting out bored from interested students).*

Alternately, the qualifying variable may be a secondary or ancillary experimental variable, which we can vary by means of procedures introduced within an experiment: for example, we might hypothesize that the superiority of a branching over a linear program will be greater for a lengthy program (or late in a session, after considerable prior study) than for a short program (or early in a period). As another example, the hypothesis might be that the superiority of a color film will be less if students know they will be quizzed frequently than if they do not expect to be quizzed. In these cases the qualifying variable can, at least potentially, be manipulated experimentally (through varying the length or placement of the program, or the kind of instructions or announcements that are made to students, etc.). In either case--that is, whether the qualifying variable is a "population" (learner) variable (as in the first example), or may be treated as an ancillary experimental variable--the form of the contingent hypothesis is the same: The effect ($Y_1$) of a primary experimental-variable treatment ($X_1$)--as compared with the corresponding effect ($Y_2$) of a comparison treatment ($X_2$)--will be greater under condition A than it will be under condition B. Using symbols to provide a convenient shorthand, $Y_1 - Y_2$ is greater if A than if B, or $(Y_1 - Y_2)_A > (Y_1 - Y_2)_B$; or, $(Y_1 - Y_2)_A - (Y_1 - Y_2)_B > 0$. We thus see that, for any contingent hypothesis, we have to deal (at a minimum[*]) with a "second-order difference," or "difference-between differences."[**]

Arguments concerning the appropriateness of simple unqualified hypotheses, versus the simpler sorts of qualified or contingent hypotheses (and versus highly complex designs in which the interaction of a multiplicity of factors might be examined) are considered in Chapter 12. For the present we merely note the fact that at least the simpler sorts of contingent or qualified hypotheses will be encountered in a number of instructional-media experiments.

# Chapter 6

# Design Steps in Controlling Alternate Hypotheses

In Chapter 5 we defined scientific method as a means for choosing among alternate hypotheses, and we have tried to establish the method of experiment as an especially good logical framework for such comparison which is characterized by control of alternate hypotheses. In this chapter, we address ourselves to some questions about what can be said concerning the nature of alternate hypotheses which should be controlled against in instructional-media experiments.

In each experiment, all the propositions that may be used as arguments which might weaken the logical rigor of the process through which we can attribute the observed effects to our experimental variables are the basis for alternate hypotheses. But also, only those which might damage the logical structure or are reasonable sources of bias represent alternate hypotheses to be considered in that particular experiment. There is no single set of alternate hypotheses which will apply to every experiment. We here re-emphasize the idea which is repeated throughout this book--namely, that the most important aspect of experiment is its logic. And we keep repeating this statement even though it does not give the prospective experimenter much help in assuring that the experiment is logically tight or rigorous--or possibly because we can hardly do more than affirm this contention.

We may say that there are two alternate hypotheses which must be taken into account in every experiment on the effects of instructional media. These are:

(1) The effects observed can be explained in terms of selection bias --that is, to systematic pre-existing differences among the people who are the subjects for the several experimental treatments; and (2) The Null Hypothesis that the effects observed can be explained in terms of the operation of chance factors; thus, concerning the null-hypothesis Fisher says "Every experiment may be said to exist only in order to give the facts the chance of disproving the null hypothesis." (Fisher, 1960, Sec. 8).

Beyond these two, the kinds of alternate hypotheses discussed in this chapter are called "recurrent" rather than "standard" alternate hypotheses. That is, they are not standard alternate hypotheses in the sense that, without control against them, any experiment on instructional media is necessarily invalid; but they do occur with sufficient frequency that they merit listing. In the main, our listing will be of classes or types of such hypotheses, rather than specific examples. This will constitute a check list which any experimenter might well run through to assure himself that he has provided for adequate control where appropriate.

Designing an experiment

We would now like to outline a set of steps for designing an experiment which will form a context for considering alternate hypotheses and their control. The procedure to be outlined is a procedure, not the procedure. Not all experimenters (and not all really good experimenters) go through all these steps in this order. Each experimenter will develop his own work habits, but the operations and the purposes of the several steps should be accounted for in some form. The experienced researcher would feel hobbled in taking such disjunctive and relatively small steps, and the order, too, might be strange and uncomfortable. In the main, the experienced researcher handles these steps implicitly, rather than explicitly. He probably will not become explicit until time for a final review before starting to collect data. It should be noted, further, that not all seasoned experimenters will develop the same experimental design for testing a given experimental hypothesis.

From our earlier discussions of the nature of experiments, we can say that an experiment is devised to test (1) whether the effects predicted by the hypothesis can actually be observed empirically, under conditions where (2) other possible explanations for the effects observed are controlled against. We might recommend that a design be established first to take care of (1), temporarily neglecting (2). That is, the first step should consider the experimental hypothesis only, neglecting any alternate hypotheses--neglecting even the two standard alternate hypotheses.

The word "neglecting" is deliberately chosen over such possible words as "avoiding." To expand on previous observation on the work habits of the experienced researcher, we suspect that if you are not a complete novice in the design of experiments, instructional-media you will have difficulty even thinking about design without taking the two standard alternate hypotheses into consideration. In fact, as you become more experienced, you will automatically (not necessarily explicitly) be considering more and more alternate hypotheses in the first cut. This should not be avoided if it "comes naturally." Our recommendation to neglect means to develop the first model of the experimental design without deliberate concern for alternate hypotheses (with the possible exception of the two standard alternates.)

Another way of stating the recommendation would be: First, develop an experimental design to test the experimental hypothesis, assuming that all alternate hypotheses will be controlled against automatically. (You will, indeed, find that many are.) Then, test the assumption and make corrections as necessary. Each time you make a change in the design, the assumption of adequate control must be re-evaluated. The most experienced experimenter(s) needs to make this kind of successive correction. The advantage of experience is that the recycling process of successive approximation to a satisfactory design can be shortened.

### Step One: Selecting the essential elements of design

For the first step, then, considering only the comparisons stated in the experimental hypothesis, what kind of situation will be a logical test of the effects predicted? The problem, of course, is to design and/or select a situation which will permit logically sound statements about the hypothesis. At a minimum, selection of the situation will involve at least the preliminary definition and specification of

(1) a learning task or tasks
(2) the instructional objectives
(3) the kinds of subjects (learners) to be taught
(4) the selection of the subjects
(5) the instructional instrument(s) or program(s)
(6) the associated utilization procedures and the variants thereof to be introduced
(7) the means for observing and/or measuring effects
(8) the conditions under which the measuring instruments will be used.

As the designing proceeds, you will sharpen and may change these definitions and specifications, but the job should be done as well as possible at this first stage. One might suspect that the definition of the test situation would emerge from the logical elaboration of the implications of the hypothesis. Indeed, we said as much in Chapter 5. We did not say, however, that there is any automatic, or even describable process for doing this. The direction and incisiveness of the logical elaboration will depend on some unspecifiable creativity and skill of the researcher.

Moreover, the degree to which this elaboration can be a straight deductive process, like solving a mathematical equation, varies tremendously. Paradoxically, a straight deductive approach is most applicable at the two extremes of very well-developed theories and very pedestrian hypotheses--in the one case, because the load of hard thinking which has already gone into the development of the theory has resulted in a good logical framework; and in the other case because the very simple-mindedness of the proposition doesn't afford much opportunity for elaboration. Where can you go with the hypothesis: "My third grade students in the class I am now teaching at the George Washington Elementary School in Wheeling, West Virginia, will learn the names of more capitals of the fifty states of the U.S.A., if I give them Presentation A, which I hold in my right hand to read, than they would if I gave them Presentation B, which I hold in my left hand to read, in the next half hour?"

There is a further difficulty in the problem of accessibility, which cannot be solved by deductive inferences from the hypothesis. A situation which is not available for exploitation cannot serve to test a hypothesis, however rigorous and beautiful the logical implications for the hypothesis may be. (It could, of course, happen that you can deduce several situations and choose one on the basis of their relative accessibility.)

What usually happens is that, somehow or other, a possible test situation occurs to the researcher, who then examines its accessibility and its logical adequacy for testing the hypothesis. The procedure we propose for evaluating the logical adequacy, we will call the "Refrain." We call it the "Refrain" because this procedure (or some other which will serve the same purpose) should be repeated every step of the way throughout the design of the experiment. The Refrain will apply throughout this discussion whether it is mentioned explicitly or not.*

### The "Refrain"

As the researcher goes through the design procedure proposed in this chapter (or any other procedure), at each step he has, at least tentatively, a schema (design) for testing the experimental hypothesis. The most critical question you need to ask yourself is: "If I go ahead and use this design, what will the results permit me to say?" It is this question that we refer to as "the refrain."

Now, of course, you don't know in advance what the results will be. Your hypothesis *predicts* what the results will be--providing the design is appropriate to the test of the hypothesis. But if you *know* what the results would be you shouldn't do the experiment unless you have reasons other than testing the hypothesis. You can say something about what the results *might* be, even if it is no more than "Treatment A will result in better learning than Treatment B, or vice-versa." (There are, obviously, these two possible results of any experiment on instructional-media effects.) You should, however, try to state as many possible results as you can, and as exactly as you can. "Possible," here, should be understood in the broadest sense, without any restriction as to likelihood.

Now, pretend that you have done the experiment, and pretend that you got each of the possible results. For each of the possible results consider what the situation would be if you had done the experiment and had got that result. The question now becomes, "Given what I have done and the results I got, what logically sound statements can I make?" Really better is to ask the question in the form, "Given what I have done and the results I got, how can I explain such results---to what can I attribute the effects observed?" If you cannot make a clear attribution which will permit you to know whether you can accept or reject your experimental hypothesis for every possible result, you know that there is something wrong with your hypothesis or with your design.*

At this stage, since you have only gone through Step One, you know, even without going through the Refrain, that your design is defective. In fact, you know that you don't have an experiment because you haven't yet considered the alternate hypotheses. Nevertheless, you should go through the Refrain for what you may learn about the logical adequacy or inadequacy of the several aspects of the situation you have chosen. You are then ready to review the eight listed aspects of Step One (that is, Step One can now be "recycled") to correct deficiencies found through the Refrain. (We are assuming that the first application of the refrain has not led you to decide to throw out your hypothesis and start over again).

Use of the Refrain should not necessarily be limited to the end of each step, and partial "Refrains" should be used with some frequency. By "partial Refrain" we mean addressing the question, "How might this result be explained?" in terms of only a single possible result, or of only a few of the possible results--as differentiated from a complete run down of all the possible results which you can anticipate. We will not try to say when a partial Refrain is appropriate, but will leave the answer to that issue to emerge from your own experience. You will find that sometimes, when you make a small adjustment or correction in the design, it may be clear that such a change need be examined only with respect to one or two possible results; or, when you want to make a choice between two or more corrections for a defect, the decision can be made by examining only one or two of the pretended experimental results.

With accumulated experience in experimentation, you will make your own modification of, or abandon, our particular outline of design steps. (Whatever your state of naivete may be at this reading of the guidebook, it is assumed that you will soon have as much authority as the authors for issuing such pronouncements and injunctions.) However, we enjoin you not to abandon the Refrain, unless you replace it with an equivalent substitute. If you find that you can work more efficiently without as frequent use of the Refrain as we recommend, by all means do so. But always go through the Refrain (or its surrogate) with as much rigor and comprehensiveness as may be available to you, before you start to impose treatments on subjects.*

Step Two: Explicit consideration of the two standard alternate hypotheses

If you have not already done so in Step One, now is the time to take up the two standard alternate hypotheses. As we have said, it is very likely that you will already have at least partially accounted for those two alternate hypotheses in Step One, though you may not have deliberately or explicitly considered them. Certainly, we expect that it will not be very long into your career before you collapse Step Two into Step One.

At this point in the development of the design, we would recommend that you make a tentative selection of simple randomization as the control procedure against the learner-differences or selection-bias hypothesis, unless some other clearly feasible control procedure has immediately obvious advantages in terms of precision.** Remember that until you actually impose treatments on learners, you will have unlimited opportunities to change--points at which you can explicitly consider whether you should stay with whatever control techniques you may have chosen at this stage of the design.

The reasons for at least temporarily electing simple random assignment of subjects to treatments is that this is the easiest thing to do, and this randomization will automatically provide a valid accounting for both of the two standard alternate hypotheses. You will thus quickly have an experiment. It may or may not be an efficient experiment, but it will be (in some sense at least) a valid experiment, which you can then consider as a basis for review, evaluation, and improvement. Another justification is that, in our opinion, the basic reasons for using procedures other than simple randomization of subjects involve only the sensitiveness and efficiency of experimentation. We feel it is better to postpone explicit consideration of these two characteristics until after a valid, or reasonably likely-to-be valid, design has been established. Now, before proceeding to Step Three, invoke—for a second time—the "Refrain."

## Succeeding steps in design

Every issue that will be mentioned from here on in the outlining of the experimental-design procedure may have been taken care of before you reach our Step Three. (We might say that there is now already a non-zero probability that you have a valid design for an experiment.)

You have, presumably, done some search and reading of the relevant experimental literature, and some concerted thinking, before you even formulated the experimental hypothesis. Very likely the experimental hypothesis itself encompasses implications of control against the most important conflicting theoretical factors. You have, also, automatically controlled against many relevant alternate hypotheses—automatic in the sense that (1) your work habits are such that you have introduced many controls without explicit consideration; that (2) control considerations are forced in whenever you try to keep your thinking about an experiment logically rigorous; and that (3) almost any control for one variable will control for many others. In particular, once you introduce randomization, you have controlled against many of the alternate hypotheses you might consider. And you have been through the Refrain at least twice and corrected for deficiencies revealed thereby. Nevertheless, we still recommend that the remaining steps be taken up explicitly.

## Step Three: Examining the design of the experimental treatments

You have now made a number of provisional decisions about just what you are going to do and what you are going to have to control in the experiment. Now each of these decisions should be examined to see whether any new factors have been introduced which require new forms of control. Every time you select some aspect of implementation, new factors may possibly have been introduced which can cloud the interpretation of the results.* If everything has been taken care of already, good. But you may well find some things you have missed earlier by going through Step Three. There are many traps along the way for the designer of instructional media experiments (wary or unwary). A sizeable list of casualties can be drawn up from an examination of the literature.

A particularly dangerous trap, and the one with which this section primarily deals, lies in the selection (or construction) of the instructional instruments (programs). It is not easy to assure that the two or more instructional programs differ with respect to the hypothesized experimental variable (or variables) and only in this respect. Being able to tell that they do differ with respect to an experimental variable is largely a question of the adequacy of the definition of the variable. (Some further discussion of this question is given in Chapters 12 and 16.) The problem we are concerned with at present is whether, in introducing the experimental variations, other differences which might bias the experiment may also have been introduced.

As a primary example we shall consider at some length an experimental question to which we have referred several times previously--the question of the influence of color (as against black-and-white) as a factor in the effectiveness of an instructional film. We use this example not because we think this question is necessarily an important nor even a particularly interesting one. Rather, we use it because it serves as a convenient illustration of ways in which there may be serious gaps between the mere statement of an experimental variable, in abstract terms, and its defensible implementation in terms of the actual construction of experimental treatments.

For almost every experimental variable in instructional media that can be mentioned, similar difficulties arise, at least in principle. For example, one of the most common of the confounding factors introduced in experimental treatments is differences in instructional time that may be introduced when E is trying to investigate some other variable (e.g., response mode in a self-instructional program, or type of review sequences —cf. later discussion in Chapters 13+14). In the color/black-and-white comparison this particular confounding factor does not necessarily enter because we can presumably make corresponding film treatments that are equal in length (and, hopefully, in all other respects, except only for the difference in color). But, as we shall see, even this apparently simple form of variation is fraught with potentially perplexing difficulties when we leave the armchair where we did our hypothesizing and get into the laboratory to start the work of constructing actual experimental programs. Even though some (or all) of the difficulties to be discussed for this example may Sometimes turn out to be of little importance, they are worth examining in two respects: (1) as examples of the kinds of difficulties and ambiguities that are at least needful of being considered—if only to be dismissed as inconsequential or irrelevant; and (2) because at least until they are examined, such potential difficulties in clean, unambiguous implementation of any experimental variable can not necessarily be so dismissed.

Suppose, then, that we wish to do an experiment concerned with the comparison of black-and-white versus colored instructional films. This would seem to be a relatively easy variable to define. It is surely easy to tell whether a film is black-and-white or in color (that is, to satisfy the first requirement that the treatments are, in some manner at least, differentiated with respect to the proposed experimental variable). Maybe not good color, however (or, for that matter, good black-and-white) but, at least color, as versus black-and-white. Insuring that the instructional instruments, then, do somehow differ with respect to the experimental factor seems to pose no problem in this case. We also want to try to insure, however, that they differ in no other relevant respect. That is, we need to ask whether, in the way we go about getting the color film and the black-and-white film to compare, we may introduce some bias which would confound the experiment. Let us examine some ways we might go about getting the two films, and the implications of these ways of doing so.

It is possible to make a black-and-white print from a picture that was made in color. It is, in fact, common practice for educational film producers to photograph in color and offer the customer a choice of a color print or a black-and-white print. The black-and-white print will often be better when the color photographs are originally made with the intention in mind that black-and-white prints will be made. Even so, especially when photographing from nature or, in any case, where coloring is taken as found instead of chosen, definition and differentiation of objects, shapes, and conformations may be lost or weakened. These, and other differences, may in fact, be pertinent to our hypothesis and, if so, we will want just these differences to appear. This would certainly be so if our question was simply whether to buy the color or black-and-white print of a particular film. On the other hand, such differences might be damaging. Their relevance depends on the hypothesis being tested.

There are also at least two other ways that we might choose to get the two film versions. We might photograph simultaneously on both black-and-white and color negative. Using a single lens and a beam-splitting technique, we could present exactly the same image to both negatives. We might also choose to make the film first in black-and-white and then do it over again with color film.

The beam-splitting method may seem attractive. It can give a better black-and-white print than working from a color photograph original, since we can get to the release print in at least one "generation" (reproduction step) less. But we may have some new photographic problems. For instance, color and black-and-white have different "make up" requirements for actors. If we make up the actors for color, photograph in color, then print in black-and-white, make-up may be no problem. However, if we make up for color and photograph in black-and-white, we may get some very funny looking people.

In any case, if we set up to photograph for color, we will not get the best possible black-and-white and vice versa. What is considered good lighting practice for black-and-white is considerably different from what is considered good lighting practice for color. If we wanted to get the best black-and-white to compare with the best color, we would have to photograph independently with different conditions of lighting. But it would indeed be quite a trick, requiring great precision to say the least, to go through, say, twenty minutes of filming twice and have no differences in what was photographed. Would the differences we didn't avoid be important--need to be controlled against--or not? We may have to compromise--that is, to decide whether such differences are likely to be more important than the differences in clarity or fidelity of tone that result from an alternative technique such as printing black-and-white from a color original. Assuming that we were able to present exactly the same material to the cameras twice, there would still be differences in clarity. We must ask: are these differences in clarity of intrinsic relevance to our hypothesis--that is, are they among the very differences we want to compare, or are they differences which will confound the interpretation of our experiment? We could complicate the discussion further by bringing up the fact that the differences between the black-and-white and color will depend on the particular film emulsions we use, and even whether we photograph in 35 mm. or 16 mm., because of the different emulsions available in the two sizes.

The discussion of black-and-white and color films may or may not seem discouraging, picayune, or both; but it should point to at least two morals.

First, you must always look for the implications of the form of implementation you have chosen. You can pretty well bet that, any/time you make up any two versions of an instructional instrument, whenever you are manipulating media variables, you will probably introduce some unintended differences. The unintended differences may be of no importance, but you cannot say so unless you examine them explicitly.

Second, you should consult experts whenever you are not yourself an expert and whenever experts are available to consult. You cannot be an expert on everything. (To this advice there is a common rejoinder about who pays the experts. We can only say that you look for that money the same places you look for any other money for your research. To the graduate student who has no place to look but his own pocket, we suggest that you survey your own campus. You will find plenty of people prepared to gladly give time and guidance to students, even though they may be from outside their own departments.)

### A digression on consultants and decisions

In using consultants, try to be sure you know what you want. The consultant should not be making your experimental decisions. You may properly consider yourself to be the only expert on your experiment as such--first, because the expert may not be competent to make the decisions (remember that it is not possible for him to be an expert on everything, either); and, second, because the scientific community wants your contribution. Humility is fine, and it is wise to listen very carefully to the older hands, but if somebody else makes the experimental decisions, it is his experiment, not yours; you are then his research assistant, not an independent investigator. Serving as a research assistant is undoubtedly good training, but--sooner or later--you must decide what you want your contribution to be. You might decide that your greatest contribution can be to continue serving as a research assistant to an experienced investigator, or even that you will make your contribution to the advance of a science of instruction indirectly--say, by speculating on Wall Street and (hopefully) getting into a position to provide financial support for research. Research takes lots of man hours and lots of money; but the commodity most valued in the scientific community is good ideas. As far as the scientific community is concerned, your ideas are sought and are, potentially, as valuable as those of any Distinguished Service Professor.

You go to the consultant for only two things: (1) For specific information or skill which he may have and you do not, and (2) for exploring the possible implications of decisions you may make or actions you may take. If you are responsible for the success or failure of the experiment, it is you who must make the decisions.

## Other "morals" for the design of experimental treatments

A third moral that can be drawn from the example considered above is that there is very little that can make any difference to a poor hypothesis. For the vaguely stated "comparison of black-and-white versus color instructional films" we started with, we could not say whether any of the film production problems mentioned were important or not. (The respects in which more incisive formulation of such a hypothesis might be arrived at will be considered in a later chapter.)

As a fourth moral, we may note that careful design, and systematic examination of the design, can lead back to reconsideration of the thinking behind the experimental hypothesis and its consequent refinement and improvement—or, in some cases, its abandonment. We might, for example conclude: "Well, the way my experiment is shaping up, I can't really tell much about the effects of color in instructional films, so I will settle just for getting the best evidence I can about which of two particular films is better." Then we might decide, on further reflection, that this is not a sufficiently important question to be worth the cost of doing the experiment.

Having examined the means for implementation of an experimental variable to see whether it is confounded by the introduction of the new factors or unintended differences, and having found some, what can you do about them? These factors have the same formal status in the design of an experiment as alternate hypotheses derived in any other way; the same considerations apply to these as to any others. They may or may not be important; you may or may not even be able to determine whether they are important. If you decide they are, you may or may not be able to design a really adequate control. If you aren't you may or may not decide that the experiment is still worth doing.

There are, however, three ways in which the hypothesization of biasing effects, due to confounding of factors in implementing an experimental variable, differ from other alternate hypotheses. First, since you have been able to identify these differences, you know they are going to be in your experiment unless you do take definitive action. Unlike the general case where you want to avoid alternatives which may get in without your knowledge, you know these will be in. Second, such confounding factors in the experimental treatment are often mistakes in the sense that something dropped in a crack when you were trying to sweep the table clean, rather than representing a fault in logic. That is, your logic was correct and you knew you should clean off the table but some crumbs got away. For these you need have no special concern for fundamentally different controls; you just need go back and sweep a little better.

On the other hand, alternative factors which are revealed from examining the implementation may be especially hard to control against, or may have peculiar implications for the underlying theory. They may be of especial importance just because of the fact that you were trying diligently to select a test situation that could provide as clean or as relevant a test of your hypothesis as possible, and were not just taking a test situation out of a hat.

The peculiar question to look into for possible irrelevant factors that may confound the experimental variable is whether the factors are or are not *necessary* concomitants of implementing the experimental variable. A good way to attack this problem is to consider other ways in which the experimental variable might be implemented. If the hypothesis itself doesn't dictate the choice, consider whether the troublesome factor would still occur had you made a different choice. If the hypothesis doesn't specify the instructional medium, would the unintended factor still arise if you were to implement the experiment with another medium? Other tasks? Other subjects? Other learners?

The "other" implementations must be relevant to the hypothesis or they are of no help. They should also be situations which *could* occur in real life. That is, they must be potentially accessible to somebody sometime--now or in the future. If they are inaccessible because they can't conceivably be produced in real life--not just because, say, you can't get permission to utilize them--then they are also irrelevant.*

On the other hand, if you *can* define a test situation where the confounded factor would *not* occur, then it is, obviously, not a necessary consequence of implementing the hypothesis and should be avoided. Such situations, however, may not be available to you (for the present, at least), or for some reason they may be available, though undesirable. If you cannot devise a control, and the factor is consequential, within situations available to you, you will have difficulty justifying any decision other than to abandon the experiment. But not necessarily; for it can sometimes be the case that the sheer experience of wrestling with a problem empirically is rewarding enough to justify the effort in spite of a damaging flaw.

If, moreover, you find that the unintended factor is inextricably bound to the experimental factor, you may have a finding of considerable scientific import. If the experimental variable you have thus called into question has wide acceptance in current instructional-media theory or practice--and assuming that you have done a pretty good job (not necessarily demonstrably exhaustive) of examining the possible array of implementing situations--you may have shown that current theory is inadequate and needs rethinking.

Unfortunately, what is commonly found from such an examination is not so dramatic or satisfying. More often what is found is that the experimental variables are poorly defined. In part, this is a consequence of the fact that our field of research is too new and immature to have developed good definitions. Demonstrating the inadequacy of current definitions can, therefore, be a significant contribution to the advancement of the science. But, by "poorly defined," we really are indicating that such an exhaustive analysis as we have implied should not have been necessary, and that is it unfortunate that so much effort was invested when it needn't have been. One of the major arguments against the scientific justification of such "cross-media" experiments as the comparison of, say, film versus lecture, is in terms of their inadequate definition of variables and the neglect of this Step Three in the formulation of the research. When an experiment compares teaching some subject matter or task by film versus television, to take another example, is the comparison really between a definable factor which can be called "film" and a factor which can be called "television"? If there is, in fact, a comparison between factors which are unique to film and factors which are unique to television (rather than merely adventitious concomitants (see Chapter 2, p. 9 ), did the implementation also introduce other differences which have no necessary relationship to one presentation's being by film and the other by television? Historically, it is usually easy to tell that there are adventitious factors which differentiate the two treatments. Logically, it is almost certain that there will be adventitious factors if the unique qualities of the medium are not defined.

To name a third example, in comparisons of instruction by some instructional medium (pick any one) with "conventional" instruction, what, in terms of meaningful variables, could the definition of "conventional" be? A most common flaw of comparisons of something-versus-conventional-instruction is that the "something" is simply better prepared than the conventional instruction--prepared by more competent people making more and better use of teaching principles not necessarily at all inherent or unique to the medium. Thus at a symposium on television in military instruction, one of the participants observed that it seemed to him that the most strking difference between the instructional treatments provided by television and by other (or "conventional") military instruction was that the television presentations were prepared by ten-thousand-dollar-a-year professionals while the "other" instruction was prepared by Privates and PFC's.

## Step Four: Other experimental elements

The foregoing extended discussion, presented as "Step Three" in designing an experiment, has dealt primarily with a special class of potential alternate hypotheses about the potential causes of observed differences in instructional effects--causal factors that come into play in consequence of the way in which the experimental variables may have been defined and implemented. It has seemed desirable to consider these at some length because of their typically crucial bearing on the successfulness or unsuccessfulness of instructional-media experiments. Further discussion of related considerations will be taken up again in Chapters 12+15 (cf. also Chapter 7 in relation to classification of independent variables).

Before proceeding to our suggested further steps we make two remarks. The first is that we would introduce here, as "Step Four" in design, a parallel examination of the implications of the way in which the dependent variables in an experiment are to be implemented--that is, the way in which "tests," or other measures of effects are to be designed in relation to the logical implication of the hypothesis. We shall, however, defer consideration of this aspect of design and its implications to a separate chapter. (See Ch. 8.)

The other remark can perhaps, by now, be predicted by the reader. It is, of course, that at the end of Step Three and Step Four there is need once more to consider:

The Refrain: What, if I now proceed with the experiment as designed, will I be able to conclude from the possible results I might obtain? And if what I could conclude is unsatisfactory, what further adjustments or changes in what I have tentatively decided, on the basis of Steps One through Four, are now indicated?

Step Five: Enumeration and weighing of alternate hypotheses

We now recommend that the researcher write down explicitly all alternate hypotheses--both those thus far considered and any other ones he can think of. Force this list: that is, make it as long as you can, without trying to make any but the grossest judgment about relevance or importance of the alternate hypotheses as yet. Put down everything that could be considered to have any relationship at all to the possible outcomes of the experiment. You will, of course, have to make some more or less arbitrary delineations; if you don't, you will spend the rest of your life on this step. In general, try to guide your listing by consideration of those variables which can be considered or have generally been found to be related to the control or description of learning; and allow yourself to be just a little bit silly in listing some hypotheses which are just over the border of sensibleness.

By Step Five you have already, explicitly or implicitly, made quite a list of alternate hypotheses from your literature review and from exercising the preceding steps. Now make the list explicit and try to push it to completeness. Given the list, for each item make an explicit judgment of importance to the expected possible experimental outcomes. For each item, consider the consequences to the logic of the experiment if that hypothesis were not controlled against. On this basis you can render a judgment as to whether you would classify each one as a relevant or irrelevant alternate hypothesis.

For each relevant alternate hypothesis, you must, of course, choose a technique for controlling. What you may do about hypotheses which are not or are not clearly in the relevant category cannot be stated so directly. You might select one of several courses of action.

You may decide to disregard them. It is your experiment, and you may decide whether you want to control against some alternate hypothesis or not. (If you think control is unnecessary and somebody else thinks otherwise, let him do his own experiment with whatever controls he thinks appropriate--unless perhaps "he" is a sponsor or a member of your thesis committee.) This decision should be on conviction, and not as cavalier as the preceding may sound. If a given alternate hypothesis is easy to control against, you might as well control. Actually, you will encounter very few of these hypotheses that some other control procedure you will be using anyway will not take care of automatically. Proper randomization will control against an overwhelming majority of all possible alternative hypotheses. The only argument against using randomization as the control technique of choice, where it will work, is that you may increase the experimental error thus reducing the precision, and thus reduce the likelihood of being able to reject the null hypothesis when you, if truth were known, should (the so-called "Type II error"). If you are right in concluding that a given alternate variable can have only an inconsequential effect on the dependent variables, there will be no effect on experimental error (or on your experimental outcome in any other respect).

The argument against introducing unnecessary controls is that it is wasteful of resources and it can dirty up--that is, complicate--a clean design. Nevertheless, you should be willing to buy some extra insurance against having an invalid experiment, and, for personal if not scientific reasons, it is just as well, when feasible, to obviate unnecessary criticism, incompetent though it may be. Opinions differ, however, on whether you should err on the side of overcontrolling rather than undercontrolling.*

Our seemingly invidious parenthetical comment about sponsors and thesis committee notwithstanding, the researcher, whether novice or not, is well advised to take the opinions and hunches of the old hands seriously. In the absence of clear, statable reasons to the contrary, there is generally a reasonable likelihood that there may be real substance underlying these opinions and hunches. This is not reverence for the ideas of established authority which, as such, has no status in science. Rather, it is a statement of the authors' impression of their own experience. The elder statesmen may be wrong; if they were always right there wouldn't be much need for the rest of us. But even when their thinking goes off in the wrong direction, there is a good chance that there are important issues imbedded in the error. The same applies to the hunches and opinions of your contemporary colleagues. In any case, it is always a good idea to try out your design, and, in particular, your tentative conclusions about the relevance of potentially contaminating variables, on colleagues, as a "sounding board." Often, an alert researcher not immersed in the particular problem with which you are involved may have the detachment to spot weak points which you have overlooked.

Sometimes the consideration of alternate hypotheses will uncover a question which should be settled by research prior to the experiment you planned to do, or which suggests an experiment you would rather do. The best advice then would seem to be to put the originally planned experiment on the shelf for a while.

When the consequences of an alternate independent variable are in doubt in your own thinking, or a subject of extensive controversy in the field, there is a great temptation to expand the originally planned experiment to try to get some purchase on the additional hypothesis at the same time. In spite of the salient arguments about the efficient use of resources in multi-factor experiments, we feel that they should best be avoided until instructional-media research is more mature (see further discussion in Chapter 15). However, if the hypothesis is that the factor will have a contingent effect, you must go to a multi-factor design or, at least, a design which can show second-order differences.

Writing alternate hypotheses out explicitly and making individual, explicit decisions about control has real value for improving the experiment being planned, and also for future experiments the researcher may undertake. Even though the experimenter, especially the experienced one, may already have taken care of all the required control by the time he gets to our Step Five, something like this step should be followed. If the work has already been done (just as we said for other steps), no great burden of new work is required. But we believe that decision making is more effective when the considerations are made explicit. If nothing more, the explicit procedure will serve as a check on decisions which may have been made. We also believe that this procedure will

[* Refers to a final portion of Chapter 10 which is not included in the present dittoed version.]

contribute to the clarification of the underlying theory by explicitly pointing to possible implications and ramifications.

A useful aid to the listing of alternate hypotheses is a classification of "Factors relative to the validity of experiments in social settings" described by Campbell (1957) and by Campbell and Stanley (1962). It is useful to think of these "factors" as categories of recurrent alternate hypotheses which are frequently relevant, but not as "standard" alternates which definitely apply to all experiments.* As a matter of fact, some of the factors they list do not apply in any true experiment, but are applicable only in quasi-experiments (see Chapter 10)* in which the random-assignment requirements for a true experiment are not met. However, others of the factors they list may be relevant in any instructional experiment, and it seems worthwhile to present their complete list.

Campbell and Stanley (1962) list twelve categories, or classes of extraneous variables. The factors they list are the following:

(1) "History": the specific events occurring between the first and second measurement in addition to the experimental variable.

(2) "Maturation": processes within the respondents operating as a function of the passage of time per se (not specific to particular extended events), including growing older, growing hungrier, growing more tired, and the like.

(3) "Testing": the effects of taking a test upon the scores of a second testing. (These may be facilitative, inhibitory, and/or "interactive.")

[* Refers to a final portion of Chapter 10 which is not included in the present dittoed version.]

(4) "Instrumentation": changes in the calibration of a measuring instrument or, particularly in instructional data, changes in the observers or scorers used, which may produce changes in the obtained measurements.

(5) Statistical regression: a predictable tendency to regress to less extreme scores when groups have been selected on the basis of initial non-average status.

(6) Biases resulting in differential selection of subjects for the comparison groups.

(7) Experimental "mortality": differential loss of subjects from the comparison groups.

(8) "Selection-maturation interaction": which in certain multiple-group quasi-experimental designs (see Chapter 10, p.    ) may be confounded with the effect of the experimental variable.

(9) "Reactive" or interaction effect of testing: the possibility that giving a pre-treatment test may increase or decrease the respondent's sensitivity or responsiveness to the experimental treatment, and thus make the results unrepresentative of the effects of the same experimental variable for the unpretested population from which the experimental subjects were selected.

(10) Interaction effects of selection biases and the experimental variable.

(11) "Reactive effects" of experimental arrangements, which would preclude generalization about the effect of the experimental variable upon persons being exposed to it in nonexperimental settings.

(12) Multiple-treatment interference: a problem that arises when multiple treatments are applied to the same subjects and a particular problem for certain kinds of one-group experimental patterns.

The first eight of these categories may be considered as primarily relevant to the "internal" validity of an experiment. That is, if these factors are not controlled in the experimental design, they can produce effects which would be confounded with the effect of experimental treatments (in our case, the instructional programs that one is investigating). Since "internal validity" necessarily implies reference to structural and procedural aspects of an experiment's design, these factors are differentially applicable to different types of experimental structure and procedural pattern (cf. Chapter 9, p. ).

For example, the first two factors, history and maturation, will obviously apply only to a "Before/After" pattern (see Chapter 9) in which both pre-treatment and post-treatment tests are given. The "history" factor refers to extrinsic or extraneous events that may be operating on the subjects between a first observation or test (or whatever the measure is) and a second. Obviously, if such "history" factors are confounded with the experimental variables, the results will be generally uninterpretable. Campbell and Stanley distinguish "maturation," as separate from "history," to refer to processes which also occur between a before and an after measure, but which operate as a function of the passage of time without reference to external events. This category includes such things as aging, becoming fatigued, getting hungry, etc. Aging, in itself, is of course usually not a problem in the typically brief instructional experiment; hunger and fatigue can usually be controlled by various means, such as breaking a long experimental program into brief sessions.

The next three categories apply to experiments using more than one test on the same subjects, either by giving a pre-test and a post-test, or an immediate retention test and a delayed retention test. The first of these categories is "testing." A subject may learn something from taking a test a first time; hence, especially unless a parallel test form is used (see Chapter 8), his score may be better the second time without any effect of the experimental treatment. In addition, so-called reactive factors may enter; that is, the measurement procedure may make the subject self-conscious or aware of the fact that an experiment is in progress. The modifications thus produced introduce irrelevant variables which will in turn affect the scores on the second test. This influence may be either facilitative or inhibitory, and may differentially interact with different experimental treatments. Any measurement process which is not part of the normal environment of the subject is likely to create such "reactive" effects, which are discussed further in Chapters 8 and 9.

The factor instrumentation involves changes in the measuring instrument over time. In a physical-science experiment, these changes would often be those in the calibration of a measuring instrument; in instructional experiments, the source of error is more often in the standards or practices of observers or raters, who may become more experienced, more fatigued, etc.

Statistical regression is a factor which always operates in a test-retest situation, but which is usually a source of serious concern only in experimental analyses in which a subgroup of subjects has been selected for investigation on the basis of deviant (non-average) scores on a pre-treatment test or some other prior measure that would be correlated with such a test. In this case, there will always be shifts toward the mean, on a post-test, that arise inevitably as a result of the imperfect realiability of the measuring instrument. Accordingly, any experimental hypothesis which differentially predicts an effect of moving initially low scorers upward or initially high scorers downward is making the same prediction, as far as direction, as will inevitably occur anyway, and therefore a control group analyzed similarly is absolutely required in order to assess whether the observed shifts were more than could be expected on the basis of regression. This will always be true regardless of the degree to which the requirements of randomization have been adhered to in forming the main experimental groups that are being compared; the artifact of regression is a special kind of selection bias which will arise whenever any subgroup is selected for differential analysis on the basis of initial level or any measure on the criterion variables or on the basis of any measure in which the errors of measurement would be correlated with those of a measure of initial level. (See further discussion in Chapters 10 and 17, and in Appendix D of Hovland, Lumsdaine, and Sheffield, 1949.)

Campbell and Stanley's sixth category--selection--falls into the class of factors that would generally enter (except for the case of selection into subgroups for differential analysis, just discussed in relation to regression) only into a quasi-experiment (see Chapter 10) in which randomization was not properly employed. Results may in this case of course be invalidated by biased selection of subjects for the comparison groups (see previous discussion in Chapter 3).

The seventh factor listed by Campbell and Stanley as a threat to internal validity is experimental mortality, or biased dropout or loss of subjects from the comparison groups. Loss of subjects who are ill (or malingering), or drop out of school, or for other reasons are not present during later phases of an experiment can destroy the validity of even the most careful randomization procedure, since the sample will no longer be unbiased. The importance of this kind of mortality will depend on how large the mortality rate is compared to the total number of subjects, and how important these type of subjects were to the experimental hypothesis. Obviously, in a captive-audience situation with a short-term or one-shot experimental procedure, the problem of mortality does not normally arise. In longer-term experiments it can be a very serious problem.' (See further discussion in Chapter 9.)

Another set of factors which, also, applies primarily to certain multiple-group quasi-experimental designs is selection-maturation interaction, and other interactions with selection, such as selection-history and selection-testing interactions. In such cases, these interactions are confounded with the effect of the independent variable being studied.

The other four factors listed by Campbell and Stanley are discussed by them primarily in relation to "external validity," or representativeness. The first of these categories is the reactive or interaction effect of testing, in which a pre-test might have a "sensitizing" or "reactive" influence (see number 3, above) and thus make results obtained for a pre-tested group of subjects unrepresentative of what the effects of the experimental variable would be for an unpretested population from which the experimental subjects were selected. (See Chapter 9, p. .)

The next category, interaction effects of the selection biases and the experimental variables, is another one which is relevant only in quasi-experiments. Category 11, *reactive effects of experimental arrangements*, may call for re-examination of the experimenter's experimental hypothesis and require a hypothesis which is much more specific than the one the experimenter originally had in mind, since these effects would "preclude generalizations about the effect of the experimental variable upon persons being exposed to it in nonexperimental settings." (Campbell, 1957, p. .) This category includes the effects of conditions which make the experimental situation artificial, such as the so-called "Hawthorne effect." For instance, knowing that he is part of an experiment, a person may either develop negative attitudes of resentment at being "guinea-pigged" or he may develop an over-cooperative responsiveness, neither of which would uniformly occur in an everyday learning situation. (See further discussion in Chapters 8 and 9.)

The last of Campbell and Stanley's categories is *multiple-treatment interference*, which may become a problem wherever multiple-treatments are applied to the same subjects in different orders. If the experiment is in fact concerned mainly with the effects of these orders or combinations, this is not a source of concern. But if one wishes to generalize to the effect of each experimental treatment occurring alone, the confounding interactions among the treatments may invalidate the conclusions.

After all relevant hypotheses that the experimenter can think of have been listed, including those suggested by the above categories as well as others, and after revisions in the design have been made on the basis of this examination, the experimenter will want to look at the resulting modified design in terms of the "Refrain": If I now go ahead as planned, what will I be able to conclude? Then we are ready to go on to the next step.

## Step Six: Calling in an expert and re-examining the design

After you have finally written down all the alternate hypotheses you can think of and discarded the inapplicable or highly improbable ones (once you have made sure that they really aren't as silly as they seem at first glance), you will want to re-examine the design for precision. Some of the problems now facing you may be alleviated by some expert advice from a good statistician on structural features of design and their implications for the size of experimental error entailed. In any case, in this connection, you will want to keep in mind the factors affecting precision in an experiment which are discussed in Chapter 11.

## Final "Refrain"

You now have designed your experiment. Before you go ahead and carry it out you should now look at the implications of the "refrain" more fully. Is this what you really wanted to be able to say? Here it would be an excellent idea to translate the refrain into more tangible terms by actually writing out, if you can muster the self-discipline to do so, a draft of your report before you have done the experiment. You certainly can write up the first parts of your report--background material, statement of the hypothesis (assuming you still have one), subjects, experimental procedures, structural aspects of design, and measures to be used. You could well even go so far as to invent some dummy data and see what problems would be encountered in making the appropriate statistical analyses. Or, instead of such a "dry-run," you could use some pilot data, if it has been feasible to run a pilot experiment. (See later discussion on pilot experiments in Chapter 11.) Then you would go ahead and write up tentative results from the pilot data, or even hypothetical results from the data you have plugged in off the top of your head. You may be surprised at how many loopholes can be filled up and how many errors can be avoided by such a procedure. Those who have tried it report that they are likely to want to make changes in the design as a result.

It is unrealistic to expect that experimenters will always go this far in a paper tryout of their experimental designs. With an unlimited amount of funds, time, and available subjects, one might perhaps prefer to go ahead and make his mistakes with a bang. But the cautious procedure just described is especially recommended in cases where the cost of the actual experiment is great, or where the range of experimental situations or supply of subjects available is limited so that it may be impossible to start over if a serious flaw in design is discovered after the available resources have been committed.

The problems encountered in analyzing data cannot all be forseen even in a complete dry-run, working out all of the steps of analysis with dummy data; certainly they are often not forseen without this step. The experimenter who has to come to the statistician for help after his experimental work has been done and his resources of subjects and experimental situations have been expended runs a high risk of being too late. Statistical ingenuity may then possibly be able to provide a palliative; but the proverbial ounce of prevention is nowhere more applicable than in the case of seeing to it that the experimental design permits a clean, unambiguous analysis of the date before the data are obtained.

# Chapter 7

# The Grammar of Instructional-Media Experiments

## Of definitions and distinctions: concerning variables, qualities and relationships

Many of the perplexing problems that arise in experimentation on educational media (as well as other problems in the field of education) can be clarified only by the results of experimentation--that is, the gradual winnowing out of truth and error which experimental results can provide, and the emergence of new concepts, hypotheses and theories, either as they are selected by the process of experimental verification or refutation, or as they are stimulated by insights to which the conduct of experiments gives rise.

At the same time, some contribution can also be made simply by clarification of terms, particularly by making clear distinctions between meanings which are a source of confusion in the planning, execution and interpretation of experiments Some of these distinctions can be readily apprehended by exemplification in concrete instances (e.g., in such experiments as those described in Chapter 4); others can berhaps best be clarified by formal definition. Relatively unambiguous distinctions facilitate efficient intercommunication not only among research scientists, but also between the research worker and the user or sponsor of research. It is useful to see basic similarities which have been masked by being described with dissimilar terms, and also to perceive important distinctions which have been beclouded through being described by like or similar terms.

The purpose of this chapter is to try to make some further distinctions which may foster clarification of thought and facilitate subsequent communication. Most of these distinctions follow more or less accepted terminology with respect to the logic, design, conduct and interpretation of experiments. However, in some instances, it has seemed necessary to introduce new distinctions or to redefine terms to suit the purpose at hand. The distinctions are not all parallel--some are more widely applicable or more important than others. Doubtless also some are better thought out than others, and hence may prove more useful over a period of time. Clear definitions are needed for the *elements* that enter into an experiment as well as for *concepts* and experimental *operations*, and for *qualities* which result from these operations.

Among the *elements* in an experiment we need to distinguish such terms as independent variable, dependent variable, learners, task, treatment, error, etc. Among the *concepts* with which we have to deal are such notions as representativeness, bias, chance, generalizability and validity. The latter, generalizability and *validity*, also refer to *qualities* of an experiment, which in trying to attain we concern ourselves with those aspects of design which involve specific *operations*, such as randomization and procedural control. We have also distinguished other qualities which result from the kinds of operations we perform (as well as from the nature of the subject matter and learners with which we are dealing) such as the *sensitiveness*, *precision*, and *efficiency* of an experiment. (See definitions given in Chapter 3.)

Table 7-1 may assist in identifying the way in which some concepts similar in terminology and meaning apply to the elements, operations and resulting qualities of an experiment.

Table 7-1A about here

Some special problems and features of instructional experiments

One of the reasons, previously alluded to (cf. Chapter 1), for the need for special consideration of methodological problems in the design of instructional experiments, as distinct from more general treatments of experimental design (e.g., Fisher, 1960. Cochran and Cox, 1950), lies in some special problems that are either peculiar to or become especially troublesome in experimenting on instruction. Some of these problems can be appreciated by considering some parallels and differences in the basic terms that apply in describing instructional-media experiments as compared with experiments in the physical or biological sciences. Differences between the two may be perceived as arising from the nature and the role of the independent, mediating, and dependent variables involved in learning (as the basic process as opposed to growth, physical change), and of the kind of relationships it is possible and/or relevant to hypothesize among instructional variables.

Table 7-3A

Applicability of Concepts to Elements, to Experimental Operations, and to Qualities or Characteristics of an Experiment

| Concept; and/or qualities of an experiment which it reflects | Application to elements of an experiment | Application to experimental operations |
|---|---|---|
| Importance | Potency (of independent variable) | Theoretical definition and/or reproducibility of independent variable; vs. confounding of related independent variables |
| | Importance of outcome (dependent variable) | |
| Systematic Bias;<br>Lack of internal validity | — | Non-random selection;<br>lack of control of extraneous variables |
| Internal validity | | Randomization; |
| Lack of generalizability or of external validity | Confounding of variables | Control (generic) |
| External validity or generalizability | Heterogeneity or breadth of sampling of Ss, tasks, and conditions | Representative sampling,<br>Replication |
| Sensitiveness | Efficacy of independent variable | "Pre-test" tryout, or pilot experiment |
| | Homogeneity of Ss and conditions | Stratification, matching, pairing, replication |
| Precision | Sensitivity of measuring instrument (dependent variable) | |

## Parallels and differences between instructional and agricultural experiments

Fisher's pioneering discussion of the design of experiments (as well as many subsequent treatments) evolved primarily in the context of agricultural experimentation. For this reason, it may be instructive to attempt to identify some parallels and differences between the independent (or treatment) variables, the dependent variables and the other factors in instructional-media experiments, as compared with those in agricultural experiments.

The problems of generalizability and external validity seem to us to offer much more difficult problems in instructional experimentation, at the present time at least, than in agricultural or biological experiments (not to mention physical experiments). Basically, the difficulty is in defining independent variables unequivocally and in measuring dependent variables (outcomes) with adequate reliability and validity. In agricultural experiments the independent variable is usually rather easy to define. It is ordinarily a homogeneous treatment which can be defined in terms of physical porperties and measured in terms of physical units--pounds of potassium sulfate fertilizer or amount of water per unit area per day on irrigated cropland, for example. By contrast, the independent variables in instruction are much more fuzzily defined, in terms of such things as "difficulty," content density," "meaningfulness" and similar characteristics that are far from having unequivocal meanings.

Nevertheless, the analogy between treatment in the two kinds of experiments is close, the instructional-media program corresponding in a sense to the watering or fertilization, if you like, in the agricultural experiment. Similarly, the correspondence in terms of the dependent variable (putting aside the additional complexities of reliability and validity of measurement in the case of the instructional experiment) is similarly easy to identify. The dependent variable in the instructional experiment is the score on some kind of a test, and may be conceived to correspond to the yield (e.g., bushels per acre or per plant, etc., in the agricultural experiment). So far as we are able to draw a parallel, the subjects or learners in an instructional experiment might correspond to the plants in an agricultural experiment. It is to the subjects or learners in the one case and to the plants in the other that the experimental treatment is applied, and it is they that produce the yield or generate the score. This parallel may be strengthened by observing that neither the growth of plants nor the responses of subjects can be manipulated directly as such; rather it is a treatment that is manipulated. The subjects (or plants) can be segregated only on the basis of selection or assignment--either assignment of the treatment to a given plant or subject or, looked at the other way, assignment of a given plant or subject to receive a given treatment.

The analogy in agriculture to task or subject matter in instruction is not so easy to see. Task or subject matter is a central element which obviously enters importantly into all instructional experiments--that is, the definition of what is to be taught-and-learned. The taught-and-learned hyphenation makes it clear that subject matter is a characteristic intimately related to both the independent and dependent variables. That is, differences in treatment are applied to a subject matter by varying the subject matter presentation, but also the choice of subject matter clearly determines the choice of outcome measure or dependent variable. That we can find no satisfactory parallel for "subject matter" in the agricultural experiment could possibly mean that here is a source of particular conceptual difficulty in discussing the taxonomy and design of instructional experiments, since in agricultrual experiments, which are the major source of conceptualization for experimental design, a paraellel element does not seem to enter in

One way of identifying a possible parallel is to focus on instructional outcomes rather than "subject matter," and to observe that in both agricultural and instructional experiments we may be interested in various outcomes. Thus in horticulture we might be interested in the size or quality of the fruit or grain, or the color and shape of the flower; and different kinds of treatment variation might well be appropriate depending on the kind of outcome sought. Perhaps the lack of an instructive parallel in agricultural experiments for the linked factors of subject matter and kind of outcome in instructional experiments arises from the main preoccupation in agricultural experiments, in any one situation at least, with a single kind of outcome--e.g., quantitative yield of a given kind. To push the analogy further, we would have to consider an agriculturalist who was interested in plants that produced a number of kinds of fruit (as in instruction we may be interested in a number of kinds of outcomes--e.g., learning of music, language and mathematics), and who, furthermore, applied successive treatments to produce each. In any case, the absence of a parallel to guide us in the design of instructional experiments seems clearly evident.

Another special aspect of educational experiments which adds to their difficulty is seen in the need to introduce the term "efficacy" of independent variables as distinguished from their potency. By potency we can mean, in common with all experiments in physical and biological science, how effective, in truth, is the independent variable postulated by the experimental hypothesis. By efficacy, in contrast, we mean how effectively a hypothesized variable has been implemented in an instructional experiment.

The need to make the distinction between efficacy and potency arises from the larger gap, in most educational and many psychological experiments, as compared with physical or biological experiments, between the conception of the independent variable and its execution. In a physical or biological experiment, we can generally specify the independent variable and define it in essentailly operational terms. In education and educational psychology we have not achieved this degree of sophistication. Consequently, we have, if we are to talk in terms meaningful to our educational colleagues, to deal with variables at least at the conceptual level which are very far from specifying unique operations. Such variables as "student participation," "reinforcement," "repetition," and "difficulty" are examples. Quite aside from shades of meaning and differences of a qualitative sort in the kind of treatments, attempts to implement a variable often may not be done "with a green thumb" or in such a way that the actual potential potency of the variable is implemented in terms of its presentation in the experimental treatment in an efficacious manner. Another factor that confuses the clear-cut definition of experimental variables is the greater tendency in educational experiments, as compared with many physical and biological experiments, for the process of measurement to interfere or intereact with the experimental variable. A good example of this is found in the comparison of the so-called "prompting" and confirmation conditions, in the case of the experiments by Kendler, Cook, Kendler, 1956; Cook, 1961; see Ch. )

The need to introduce the term "sensitiveness," as having a distinct meaning different from "precision" in describing an experiment, arises for reasons somewhat similar to those mentioned above with respect to the need of introducing such a term as "efficacy" of treatment. In discussions based primarily on agricultural or biological (or physical) experiments, the precision of an experiment and sensitiveness of the experiment can readily be used as more or less interchangeable terms (cf. Chapter 3 p. 43 ). If the independent variable has been uniquely defined so that the problem of the efficacy of the purported independent variable hardly arises, and if further the measurement is relatively unambiguous and involves, per se, little in the way of interpretation or little in terms of questionable validity, then the precision or sensitiveness of the experiment may be governed wholly by its size, by certain aspects of technique, and by its structure and the requirements for analysis of data which this structure generates. But in the case of educational experiments the situation is by no means this simple. Here, the sensitiveness of the experiment may often be determined less by these features, than by the efficacy with which the experimental variable has been implemented and by the sensitivity of the measuring instrument. (**)

It is true that in a physical, agricultural or other biological experiment, some of these same kinds of difficulties may arise. However, relatively speaking, the difficulties tend to be rather slight. Not only are the dependent measuures fairly clear-cut as a rule, but also the definition of the independent variable is fairly clear in operational terms and the degree of its implementation can be measured quantitatively. The amount of potassium sulfate fertilizer applied to an experimental plot, the intensity of heat or gamma radiation with which a piece of metal is bombarded, the introduction of a given number of units of vitamin A into the diet of a rat are--while not necessarily free from contanimating confounded influences--yet susceptible of operational definition and quantitative expression of the independent variable to a degree seldom possible as yet in educational experiments.

Of course, even in such relatively specifiable operations as agricultural fertilization there is room for ambiguity. For example, if the experimenter has put certain concentration of "loam" on a plot, the experimental treatment is much less well defined than where the kind of fertilizer applied has a chemical composition of more closely specifiable sort. On the other hand, in educational experiments, some treatments can be specified with a slightly greater degree of precision than is possible in other cases. For example, amount of "repetition" in a program may offer at least somewhat less difficulty in specifying than degree of "meaningfulness."

The difficulty of interpretation and of resolution of apparent consistencies or inconsistencies among reported experimental "findings" in educational-media experiments arises not only from the inherent difficulties of defining independent variables operationally (and of the related difficulties in defining dependent variables in terms having a unique, operationally reproducible meaning) but also in the failure of experimenters and the would-be interpreters of research results to *recognize* the degree of looseness and ambiguity in the terms they are employing. Hence conclusions are not infrequently drawn that a finding has been confirmed or has been non-confirmed when in fact the new findings have no necessary direct relevance to the findings which they purport to confirm or dispute —either because, although the same term was used to describe the independent variable, the treatments were operationally so different as to make comparison wholly misleading, or because the other features of the two experiments (including the dependent variable and the kinds of learners involved) were so different as to make a direct comparison inapplicable.[*]

### Applicability of characteristics to different features of an experiment

It is worth noting that certain terms or characteristics, in instructional-media experiments in particular, apply only (or primarily) to *variables*, wheras certain other terms or characteristics apply to hypothesized *relationships*.[*] For example, we may make a differentiation between the *tenability* of a hypothesis, the *validity* of an experiment and the *definability* of an independent variable. On the other hand, some terms (e.g., the term "importance") can apply either to an experiment or to a variable. (The "importance" of an experiment is determined by the importance of the kind of relationships hypothesized as well as the importance of the independent and dependent variables among which the relationship is hypothesized to exist.)

In some instances, we have made arbitrary distinctions between similar terms in order to keep them straight--for example we have used the term "sensitiveness" to apply to an experiment and "sensitivity" to apply to a measuring instrument or dependent variable. Some troublesome questions of where to discuss a particular term or concept arise from the fact that a given concept or term applies to more than one thing. For example, "replication" (see Chapter 1 , p. ) affects both *generality* and *precision*; "importance" applies to *hypotheses*, *independent* variables and, in consequence, *experiments*; "sampling" applies to both *subjects* and *subject matter*; and "representativeness" applies to the *independent variable*, the *dependent variable*, ~~and the~~ the subjects, *conditions of the experiment*, as well as the conclusions thereby generated (In some discussions, "validity" would appear to apply to both the *hypothesis* and the *experiment*; here, however, as noted above, we distinguish "tenability" in the case of the hypothesis to make this distinction.)

When possible, it would help to use different terms for different applications (e.g., "sensitiveness" for an *experiment* versus "sensitivity" for a *measuring instrument*). However, in some instances it does not seem feasible to think of a different term, at least without straining a meaning too much. When the meaning is not self-evident, hyphenated expressions may sometimes be used to insure clarity—e.g., as validity-of-measuring-instrument versus validity-of-experiment. In other instances, modifiers may be used to make distinctions, e.g., the distinction proposed by Campbell (1957) between "internal" validity (the validity of an experiment's structure) versus "external" validity (the generalizability of its findings, which is affected by different factors and relates, in part, to different aspects of the experiment).

Certain other terms apply both to learners and to dependent variables. Among these are such terms as stability, variability, heterogeneity, and homogeneity. Homogeneity and heterogeneity, and adequacy of sampling as it affects representativeness or generality, are characteristics which clearly apply both to learners and to subject matter.

A further attempt to show some of the correspondence between similar concepts or terms as they seem to apply to various aspects of an educational-media experiment at least, is presented in Table 7-1B

Table 7-1B about here

Major classes of variables in educational-media experiments

We have previously identified and briefly discussed the principal classes of variables in an educational experiment: (1) the independent variable or treatment variable; (2) the subject-matter or task variables (to the extent that these can be distinguished from treatment variables, on the one hand, and from dependent- variable outcomes sought, on the other hand); (3) the learners and variables associated with them; and (4) the dependent variables that represent our measure of outcome or effect. In the following sections we want to make some further distinctions and comments concerning each of these classes of variables, and particularly the independent variables in an experiment. Chapter 8 will deal at more length with the characteristics of measurement as related to dependent variables.

As a review of some of what has been said and as a preview which may provide helpful perspective for our further discussion, some of the aspects and connotations of the terms associated with the four major classes of variables are arranged in Table 7-2 so as to attempt to show some of their interrelationships. In Table 7-2 as well as in Table 7-1A and 1B, the reader can note some of the overlap in applicability of terms on which we have previsouly commented. In this table, to which the reader may find it helpful to refer again later on, an attempt has been made to indicate some of the overlaps and interrlations of terms and concepts by use of brackets comprehending related aspects of variables.

Table 7-2 about here

The features identified in relation to independent variables will be further spelled out in what follows. With respect to learner/task variables it may be noted that, depending on the point of view of the experiment, specific learner/task variables may function either as independent ("operator") or as mediating ("operand?") variables. The role of these variables in relation to the structual aspects of experiments and to the resulting experimental characteristics of validity, generatizability, sensitiveness and precision will be further discussed in later chapters dealing with replication, sampling and other aspects of experimental structure and procedure. The characteristics of dependent variables in instructional-media experiments will be considered in relation to problems of measurement in Chapter 8.

## Independent Variables

**Qualitative and quantitative variation; variables and conditions.** The term "variable" ordinarily connotes a characteristic of quality which can, in fact, be varied--often in such a way that the variation can be measured in quantitative terms. However, we do not restrict the term "variable" to quantifiable variables in the present work. A variable is simply something that can be varied. This can include all-or-none variation in the simplest sense of presence or absence, or quantitative variation (such as, say, length of a film), or variation with respect to a qualitative difference (such as color versus black-and-white in a film), which indicates a definable difference in qualitative categories even though these do not neatly fall along some quantitative series.

A possible point of terminological ambiguity is whether the term "variable" refers to the continuum or mode of differentiation or difference--that quality or characteristic with respect to which treatments differ--or, on the other hand, to a particular treatment or condition. "Length of film" is clearly a variable, but a "long film" or "short film" represent categories or conditions with respect to this variable. For other variables, however, the name of the quality that is varied and of the particular status or condition with respect to this quality may happen to employ identical terms. Thus "color," in comparing color and black-and-white films, represents both the characteristic that is varied and one of the conditions or categories with respect to the characteristic. (Although conceivably we might talk about the characteristic as being degree of coloredness, color versus black-and-white is normally treated in an all-or-none manner, and we talk about presence or absence of color.)

*"Presence or absence" of a variable.* One further comment of a terminological sort is perhaps in order. Sometimes the "presence or absence of" a variable is spoken of. However, although we may speak of the presence or absence of a "factor," it seems preferable to avoid the rather naive-sounding locution "presence of a variable" or "absence of a variable;" rather we would speak of the presence or absence of a *condition* (with respect to a variable) rather than the "presence of" or, particularly, the "absence of" a variable. This usage reflects the concept that the variable is a characteristic which may vary over all values (including zero) or conditions (including absence). The term "factor" may in most instances be substituted for the term "variable;" it is sometimes used synonymously with "variable," and sometimes synonymously with "condition."

### Importance of independent variables

The role of independent variables is of crucial importance in any experiment, and their nature in instructional-media experiments poses many complexities. Many of the difficulties in instructional-media experimentation stem from the nature, delineation, effective implementation and effective control of independent variables. The ways in which they are identified, the question of whether they are manipulated or only selected may determine whether we are dealing with an experiment or not. Which independent variables are made the focus of experiments is a central determiner, at least once the kinds of dependent variables are settled on, of what to experiment on, or what research problem to attack. The control and assessment of effects of both identified and unidentified independent variables may be regarded as the crux of experimental design; the potency of independent variables is also crucial (even though necessarily unknown at the outset) in the importance of any experiment, and the efficacy with which E is able to implement them may make or break his experiment. Further, the manner of defining independent and particularly experimental variables--in terms of their relation to theory, especially--may largely determine the generalizability of an experiment. It would appear to be worth some effort to achieve a usable taxonomy for independent variables, so that we can at least use relatively unambiguous terms in discussing their operation and their characteristics as these relate to other features of experiments.

Several useful working definitions and distinctions have already been set forth in Chapter 3. In the present section, with consideration of some of the basic aspects of hypothesis testing behind us, we can usefully review and try to pin down some of these previous definitions more precisely for use in the remainder of the book, as well as introducing and endeavoring to define some additional terms.

## A proposed working taxonomy of independent variables

We have said that a great deal of confusion could be eliminated in the discussion of instructional effects, of the properties of media, and --in particular--of experiments to test hypotheses or assertions or positions about such effects, if closer attention is paid to the definition of independent variables. This applies in two senses--one with respect to the extent to which the independent variable is described in such a way that it is reproducible; and second, with respect to the status of the independent variable in terms of its role in the experiment and the way it is introduced into the experiment or inquiry. At present we are concerned primarily with the latter aspects.*

The taxonomic distinctions we propose to employ here are summarized in Tables 7-3A and 7-3B, which will provide a framework for the following discussion.

Table 7-3A about here

Segregated (differentiated) versus unsegregated (undifferentiated) variables. By independent variables we may mean, in the generic sense, all causal factors that could affect the outcome of an experiment; but also--as has been previously noted (Chapter 3,p. 15) --we often use the term in a more restricted sense to mean identifiable causes of variations which, potentially at least, could be segregated out by the experimenter. Differentiated, or segregated factors may include, on the side of content, both theoretical-process factors and empirical, descriptive but non-theoretically meaningful factors. They are here distinguished (left of Table 7-3A) from undifferentiated or non-segregated factors (right of Table). Let us first consider the segregated or differentiated factors which the experimenter has identified.

Differentiated or segregated variables. The distinction with respect to how those independent variables that represent not only identified but also differentiated (segregated) sources of influence are further classified is fundamentally a question of whether or not they are manipulated; that is, whether we arbitrarily introduce them or administer them to learners (in the case of an educational experiment) or--what amounts to the same thing--whether we arbitrarily assign learners to a particular treatment condition. By contrast, non-manipulated factors are taken as found and are introduced only by selection of those who already are found to be characterized by a particular attribute or characteristic (as a result of events or circumstances antecedent to the conduct of the experiment).

A basis of classification previously introduced (see Chapter 3, p. 12) which can be used to further classify differentiated or segregated independent variables is that introduced by Hovland, Lumsdaine, and Sheffield (1949). This is the classification of independent variables as media variables (those which are characteristic of the instructional medium, or program, itself); external variables or utilization variables (which are characteristic of the way in which, or situation in which, the program or media is employed); and learner variables or population variables (which are characteristic of the subjects or learners). Media or external variables can both be manipulated while segregation with respect to learner variables can be accomplished only by selection.

Media variables might include such things as the use of animation or color in a film, or the response mode provided for in a self-instructional program. Utilization variables (called by Hovland, Lumsdaine and Sheffield "external" variables) could include such factors as repetition of a film, conditions of showing it, use of self-instructional program segments in spaced versus massed sessions, etc.

Non-manipulated variables may be called, generically, selection variables. These are variables which are potential sources of influence which we can identify and with respect to which we can segregate out our learners, but which we do not manipulate in a particular experiment (and sometimes, in the nature of the case, never can manipulate directly).

Experimentally manipulatable variables thus represent causal influences which can be introduced by the experimenter (e.g., by showing a particular film to a given group of subjects arbitrarily selected), whereas a selection variable can be introduced only by choosing subjects or materials which have identified characteristics which presumably reflect pre-existing potential causes of influence (such as the intelligence of the learner or his degree of prior knowledge). However, it should be noted that in some instances factors introduced as selection variables in a particular experiment might, in a more extensive experiment, have been introduced by experimental manipulation. In the first instance, we might sort out, for example, those individuals who are initially better informed, and see how the experimental treatment differentially affects them as compared with those initially less well informed. In the second instance, we experimentally give some prior instruction to certain subgroups, out of an undifferentiated pool of subjects, so as to make them better informed, prior to introducing a further experimental treatment. Another example is motivation, which can either be a selection or sorting variable (if we differentiate individuals according to some index of their motivation as caused by unidentified predisposing factors) or can be an experimental variable (if it is introduced by introducing a motivating procedure, as it was done in the experiment of Hovland, Lumsdaine and Sheffield [1949, Ch. 9] by using announcement of a test to follow the film as a motivating procedure).

Selection variables may include not only learner variables (called population variables by Hovland, Lumsdaine and Sheffield) such as I.Q., but also task variables or subject-matter variables, such as difficulty of the material. The latter have a special status in that (cf. p7-5 , supra) they normally involve a correspondence of particular aspects of presentation with particular test items, and thus are difficult to treat purely within the framework of independent variables. Task variables may also (as indicated by the overlapping brackets in Table 7-3A) sometimes be included in the experimentally manipulated variables--for example when specific procedures are employed to try to modify the difficulty level of a program without otherwise altering its content (e.g., Gladstone, 1958), or when a number of different learning tasks or items are randomly assigned to different experimental treatments (e.g., Lumsdaine, 1958x [May and Lumsdaine, Chapter 10]).* Finally, manipulated variables may be further described as treatment or control variables, as will be further discussed presently.

*Undifferentiated or unsegregated variables.* These include both identified or *identifiable* variables (which, though potentially segregatable, remain unsegregated), and *unidentified* variables that operate without our knowledge. Such unidentified variables can either operate in random fashion, or in systematic fashion (e.g., if the unidentified variables are confounded with the independent variables which we have introduced into an experiment). In the case of the latter, there is, almost by definition, no way to get rid of the contaminating influence thus introduced, which therefore represents a systematic source of error in our experiment, and one which is undetected. This kind of error, as noted elsewhere, is likely to be uncovered only by independent replication of the experiment in some way which brings to light the operation of the previously undetected systematic biasing factor.

It is to be hoped, however, that in most experiments we have enough general knowledge about the situation to arrange the experiment, through proper randomization procedures, in such a way that unidentified variables (even though potentially they *might* have a systematic effect) will have their effect cancelled out by randomization, and thus will operate only to an extent which is reflected in the estimate of experimental error. In general, it would appear that this can be done for all extraneous sources of influence which represent or are linked to learner (population) characteristics, as well as some other sources of influence, provided the condition noted by Fisher (see Chapter 4, supra, p. 16 ) is adhered to with respect to the timing of randomization in relation to other elements in the sequence of experimental operation.*

In the case of identified variables we have more choices with respect to how they can be handled. One possibility is that a conceivable potential extraneous source of variation is simply judged in the opinion of the experimenter to be inconsequential and is ignored. For example, Campbell and Stanley (1962) suggest that in educational experiments orientation of the gravitational field might generally thus be ignored, though it could not be ignored in physical experiments. The risk that an experimenter takes in ignoring a conceivable factor which he has identified is necessarily based on his prior experience and judgment, and such sources of error, should they indeed turn out to be relevant or consequential, are also likely only to be uncovered through independent replication of the experiment.

*"Intrinsic" and "extrinsic" independent variables.* Table 3-B elaborates the classification of independent variables in terms of two major factors and their relations to each other.

*Table 7-3B about here*

The first factor which enters into this scheme of classification (horizontal dimension in Table 7-3B) we have already discussed: This is the way in which the independent variables are segregated (or isolated) or introduced into an experimental comparison. The segregation dimension for independent variables is diagramed in terms of the left-to-right differentiation at the top of the table. The second major factor, discussed below, concerns the role of the variables in relation to the experimental hypothesis. Considerable confusion may perhaps be avoided if the further distinctions we now propose are kept in mind in thinking about the effects of media and about experiments directed to the assessment of such effects.

An independent variable may either be an intrinsic variable, the influence of which the experimenter wishes to study, or an extraneous (or "extrinsic") variable, the influence of which he wishes to "control." (He needs to do the latter in order to avoid contaminating the results by erroneously ascribing results to the effect of an experimental variable when actually they resulted from the biasing influence of an extrinsic variable.) This important distinction between (a) intrinsically relevant (or hypothesized) variables--those that enter into a hypothesis as independent variables whose predicted effect is to be tested--versus (b) extraneous or extrinsic (non-hypothesized, potentially biasing or to-be-controlled) variables is represented by the vertical dimension in Table 7-3B.* Intrinsic variables enter as nominative terms in the proposition or hypothesis which is under consideration, and can be regarded as "intrinsic" regardless of whether they are introduced into the experiment by manipulation or by selection.

For the four cells of the two-by-two table that results from the combination of these two descriptive dimensions, the following more precise identifications can now be introduced: "*Experimental*" variable for intrinsic manipulated variables; "*Analytic*" variables for intrinsic-selection variables; "*Control*" (or "*Procedural-control*") variable for extrinsic manipulated variables; and *Statistical-control* variables for extrinsic selection variables.** [The terms "sorting variable(s)" or "selection variable(s)" can be used generically to include all selection variables which are in fact identified or segregated, whether they are of intrinsic interest as analytic variables, or are used merely as statistical-control variables.]

An experimental variable thus is defined as an intrinsic variable that is of intrinsic relevance to the hypothesis under consideration which also is manipulated--for example, presenting a film in an achromatic (black-and-white) version as compared with presenting it in color, or increasing the length of the film, or introducing an animation technique, or manipulating the "camera angle" in terms of "subjective or objective viewpoint." To take other examples in the case of self-instructional materials, experimental variables could include manipulating the length of frames, or manipulating the difficulty level of the expository material.***

We may be concerned with the *potency* (or strength), versus the weakness or inconsequentiality of the variables, with reference to either intrinsic or extrinsic variables. We may also be concerned (particularly with respect to experimental variables) with the "purity" of a variable versus the extent to which it is confounded or complex. These characteristics are related in turn to the generalizability and the reproducibility of the experiment (cf. Table7-1B). Note that the term "reproducibility" is applied generally to events, more particularly to experiments, still more particularly to treatments or independent variables, and most particularly of all to media variables (because of the preservability of media and thus the potentially higher degree of reproducibility of media variables).

In the schema for classifying independent variables in experiments on the basis of segregation method and of relevance to hypothesis as represented in the fourfold table in Figure 7-3B, we now need to make some further brief comments on the characteristics of the other cells. The upper right cell of Table 7-3B comprises "*Analytic*"selection variables, which are of intrinsic relevance to the hypothesis, but are not experimentally manipulated. In an experiment, such factors must always be of a kind qualifying the hypothesized influence of an experimental variable; if we merely have a hypothesis which is tested by comparison of groups that are differentiated *solely* on the basis of a selection variable (or sorting variable), we do not, by definition, have an experiment.

An ancillary (or modifying or qualifying, or contingent) independent variable in a contingent hypothesis thus is one *in terms of which* the *primary* independent variable's affect(s) are differentially predicted. Since we shall contend (Chapter 12 ) that such contingent hypotheses are generally more important to an evolving science of instructional media than are unqualified hypotheses, particular interest attaches to such ancillary or modifying variables. An example of an experiment in which a qualifying variable would enter would be, for example, that the affect of color (which we manipulate experimentally) will differ depending on the intelligence of the learner (here intelligence is an ancillary or qualifying variable), or that the relative effectiveness of two forms of persuasive argument will differ in terms of the initial opinion of the audience member, or that the relative effectiveness of two forms of instruction will differ depending on initial level of knowledge. In these cases, the ancillary or qualifying variable is a selection variable (population characteristic).

Qualifying or contingent variables may also be secondary or ancillary experimental variables which qualify or modify the hypothesized influence of an experimental variable of primary concern. For example, we might hypothesize that color will be more important in the effect of a long film than in the case of a short film; here, length can at least in principle be manipulated experimentally, and the interaction between color and length is a contingent relationship which involves two experimentally manipulated variables, rather than an experimentally manipulated and a selection variable. The terms "contingent variable" and the like could be used to include both qualifying selection variables and secondary variables experimentally manipulated in the same experiment, with the effects of the primary experimental variable hypothesized to depend jointly on both: For example, in the experiment on active versus passive review by Hovland, Lumsdaine, and Sheffield (1949, Chapter 9), the superiority of the active review was hypothesized to be greatest for Ss of low intelligence, when given no special external motivation, and least for higher ability Ss when motivated by being told that they would be tested.*

### Extrinsic variables

Extrinsic variables are in all cases potential sources of error. They are causal influences, the effects of which represent alternative sources of influence other than those specified by the hypothesis which we wish to test, and thus in all cases can be regarded as representing a nominative term in potential alternate hypotheses about the effects observed on the dependent variable. We do not necessarily care whether such variables are manipulated as "procedural control" variables (lower-left cell of Table 7-3B) or varied through selection (lower-right cell of table), since we are not interested in an inference about their causal influence but only in ruling out their possible causal influence.

As examples of procedural control variables that could often be extrinsic to an instructional-media hypothesis, we might cite such factors as time of day, the kind of projector that is used to present a film, the room in which instruction is given, controllable aspects of supervision or help given with the administration of a filmed instructional program, the personality of the teacher or experimenter who administers the film, or differences in the amount of time that is given for the administration of the criterion measuring instrument used to assess the dependent variable.

Examples of possibly segregated but nonmanipulatable extrinsic variables in an educational experiment (statistical-control variables) would be such learner characteristics as sex, region of origin, or hair color. However, while we might conceivably entertain an alternate hypothesis relating the effects of an instructional-media variable to such a population characteristic as hair color, it is obvious that this would generally not be the case. In general, the most relevant selection variables or population characteristics (both conceptually and, as it turns out, empirically) are those which can be theoretically related to the process of learning--such as I.Q. or other measures of learning ability. In special cases, however, selection variables such as sex or region of origin may be important--e.g., where the nature and effects of instruction involve factors other than simple, direct learning of facts and concepts or skills, or where, by virtue of prior differential exposure, sex becomes important because of being confounded with prior knowledge or interest.

The term "control variables" may be identified generically as all those in the lower part of Table 7-3B, or may be further particularized to the "procedural" (manipulatable) control variables lower-left cell of the table--that is, those that can be controlled procedurally. All extrinsic variables potentially are controlled either by randomization, selection-stratification, or procedural manipulation.

### Controlling the influence of extrinsic variables

*Randomization*. The basic method of control, which we have indicated always needs to be employed appropriately at some point in an experiment, is that of dealing with extrinsic variables by *randomization*. By this method, rather than trying to keep extraneous causal factors "constant" or comparable, they are removed as sources of *systematic* bias by being made to operate only in chance fashion. This should be done *in such a way that their influence affects the experimenter's estimate of error in the same way as it affects the actual errors in the experiment*. It should be noted that the choice of whether extrinsic variables are randomized or subjected to some other form of control is of consequence only to the precision, not to the validity, of the experiment. However, the way that the randomization or stratification is done, and the relationship between what is done and the estimated error, is a question which affects the internal validity of the experiment--which, in fact, is impugned to whatever degree the estimate of error from the experiment is not generated directly and uniquely by the experimental operations.

Procedural control. In addition to randomization (but not, in a true experiment, in lieu of it), extrinsic manipulatable variables can be "held constant" through procedural control or by stratification. This means essentially that the experimenter strives to keep the conditions constant for all experimental treatments (at least within one replication of the experiment). For example, if E feels the time of day at which instruction is given is likely to be related to the performance of the students, he may take pains to see that all experimental classes were tested at the same hour; or he may use an experimental room with controlled conditions of temperature, illumination and ventilation so as to preclude the possibility that these factors could be influencing the attentiveness or performance of the learners. Complete constancy, however, can never be achieved, and inadvertant, presumably minor, deviations from complete constancy which are not explicitly randomized are always a potential source of undetected error (likely to be uncovered only by independent replication of the experiment).

A special way that is sometimes applicable for dealing with identified extraneous factors is through the use of counterbalancing. An example of counterbalancing is in the control of so-called "practice effects" when subjects are used "as their own controls" by learning two tasks each under one experimental condition and two other tasks under another experimental condition. Here the two conditions, A and B, might be administered in the order A, B, B, A to control for practice effects.

Stratification. A third way of dealing with (extrinsic) variables (lower-half of Table 7-3B) is an important variant of "holding them constant," in which the variables are stratified. Here the attempt is not to hold them constant across the board, but rather to employ several categories or strata with respect to the extraneous variable, trying to hold certain conditions constant within each stratum and randomizing other factors within the strata rather than across the board. For example, experimental treatments might be assigned at random within each of several hours of the day, or under each of several identified conditions of physical surroundings, such as the administration of the experiment in a particular classroom.

Three objectives can be served by such stratification, as compared with holding conditions constant across the board (for the entire experiment). In the first place, an increased range of available experimental situations (and/or increased speed or economy in conducting the experiment) may be achieved by utilizing the larger number of experimental situations that thus may become available.* In the second place (see Chapter 10, p. and Chapter 11 , p. ), the use of stratification may increase the precision of the experiment by making it possible to assign a considerable amount of observed variation to the influence of the stratifying variable instead of letting it operate randomly and thereby increase the error term. Whether we will in fact gain in precision by stratification as compared with reliance solely on randomization depends (cf. Chapter 11, p. ) on the consequentiality or potency of the control variable. In the third place, additional information can be obtained which becomes relevant to additional or ancillary hypotheses concerning the effect of the control variable. For example, in addition to the information relevant to hypotheses about a program variable we might, by stratification, also get ancillary information about the effect of time of day or temperature conditions on learning. In addition, since the primary experimental hypothesis is now tested *under several conditions*, we also have an opportunity to check the *generality* of the experimental hypotheses under each of the several conditions of stratification with respect to the control variable. (See later discussion of replication, Chapter 11 , p.  .)

In the history of an experiment it may be perceived that an extraneous source of influence which it is decided to control through stratification might, in fact, interact with the experimental variable(s) in a theoretically meaningful fashion. In this case, a variable that was initially of interest only as a procedural-control variable utilized to increase the sensitiveness of the experiment might acquire additional interest as a factor entering into a subsidiary hypothesis--for example, the hypothesis that color in a film might be more advantageous under adverse conditions of showing that are conducive to lack of interest. The experimenter may also treat an extrinsic selection variable as a secondary focus of inquiry. For example, I.Q. may be stratified as a control measure to increase an experiment's sensitiveness--e.g., by assigning equal numbers of each of several I.Q. groups to each treatment--but, also, the influence of I.Q. may then be investigated--i.e., the interaction of treatment effects with I.Q. may be analyzed.

It can be observed that those extrinsic variables which *could* be procedurally controlled or "held constant" (or partially thus controlled) through stratification but which are, in fact, instead randomized, must enter into the estimate of error as a random source of variation. The other principal identifiable component of error variation is that due to variables which are *unidentified* sources of variation but which also enter into variability in the criterion performance (i.e., the measure of the dependent variable). Fortunately, all such sources that are linked to the identity of subjects who are randomly assigned to treatments will automatically enter into an appropriate error estimate based on the random assignment of the subjects.

To the foregoing discussion about the classification of independant variables we can evidently relate some major distinctions concerning bias, randomization, and internal validity* (cf. Chapters 3 and 11). In connection with internal validity we can also discuss one aspect of representativeness, that is to say, lack of bias in sampling or assignment.

*Sampling*. Generically the term "sampling" applies to the choice of all those elements (including subjects and tasks and programs and conditions of utilization) which are to be included in the experiment in the first place, since we never do all possible experiments on all possible people, with all possible treatments for all possible tasks to represent a given hypothesized factor (see below). However, the term is most commonly applied only to sampling of subjects and, less often, to sampling of tasks or learning materials. A useful differentiation can be made with respect to the purposes of sampling as these pertain to *representativeness* (or *external* validity), on the one hand, and as they pertain, on the other hand, to what is sometimes called comparability or what we shall call *lack of bias*, attained through arbitrary or random assignment (in order to assure *internal* validity).

As just noted, the question of *representativeness* enters into the validity, generality and, thus, importance of the experiment not only in respect to the kind of variable on which concern is most frequently raised--namely population variables--but also with respect to the representativeness of other elements in the experiment. This question, as related to the base for generality or the generalizability of findings, is further discussed in Chapter 11. For now we may merely point out that relevant points derserving of discussion are (a) the question of simultaneous sampling of learners and tasks, and (b) representativeness of conditions, and (c) the relation of Fisher's discussion of homogeneity and heterogeneity to the question of generality (see Chapter 11, p. ) and (d)--to some extent cutting across these--the question of replication. In general, the more heterogeneous the sampling of learners and other circumstances which are comprehended within the experiment, the greater basis we potentially have for defending, on explicit grounds of evidence, the generality of our results.

Learner and Task Variables

Before turning (in Chapter 8 ) to the other major class of variables that enter into cause-effect (or, also, correlational) relationships--namely, dependent variables--let us consider some distinctions that can be usefully made with respect to two other classes of variables which have a somewhat special status in the field of educational or instructional or pscyhological experimentation (as differentiated from experimentation in general). These are learner variables and task variables. As already seen, learner variables can form a class of independent variables, namely selection variables, which can enter as nominative terms, like other independent variables, into cause-effect hypotheses; but they also have a different status in that they are the operands on which the other independent variables operate--i.e., they are part of the predicate in the hypothesis also. (Further, it could be said to be a predicate "in the dative case." That is to say, an independent variable is applied to a learner in order to produce an effect on a dependent variable.)

*"Groups" in an experiment.* Since we normally deal with a number of learners or subjects (*Ss*) for each treatment in an experiment, we often have occasion to refer to experimental groups . Some terminological distinctions with respect to Groups of subjects are therefore needed.

The term *Group* as used herein, means a group of individuals--generally a group that is given a particular experimental treatment; e.g., a "control group" and/or an "experimental group." However, it also sometimes refers to a group of people who are instructed or tested in a common situation on a given occasion. In the latter sense, a "group" may consist of more than one "experimental group," (i.e., different individuals in the group may receive different treatments--at least, with some kinds of programs; see p.      ). On the other hand, a group tested and/or instructed in a particular situation may be only one element or component of the experimental (or control) "group" that receives a particular treatment--as when, say, six class groups comprise the experimental "group" that gets a particular treatment. The term group thus has two distinct meanings: first, as a *structural* element in an experiment (e.g., "experimental group"), and second as a procedural element (e.g., classroom group).

*Classroom group or "class"* (as distinguished from other common meanings of class) refers to a group of individuals who are instructued and/or tested together in a group situation or condition in instructional-media experiment. (Note that testing as a group does not necessarily imply group instruction, or vice versa.)*

We can use the term *experimental group* (or sometimes just "group" for short) to refer to *any* group of subjects who receive a particular experimental treatment or condition. "Experimental group" is also sometimes used in contradistinction to "control group." In this sense, an experiment might include one or more experimental groups and one or more control groups. More generically, however, "experimental group" refers to any group receiving any treatment in an experiment, including control groups that receive "control" treatment.**

The term control group is used to refer to groups of subjects used in an experiment to provide a control on some factor or cause that could figure in an alternate hypothesis. A more generic term would be control treatment. The general nature of control has been previously discussed in Chapter 5 . There are many forms of control which do not involve control groups (cf. Chapter 5, p. 23 ). An experiment may need no control group, or one, or several--depending on the requirements of the hypothesis under consideration and on the conditions of experimentation, or depending on the caution which the experimenter may feel is essential to safeguard the validity of the conclusion which he wishes to draw concerning the proposition which he wishes to test. As noted in Chapter 5, the term "control group" is used in a variety of ways, and in reading the literature it is necessary to make sure you know which way the term is being used. At least three meanings need to be distinguished which will be encountered in discussions of experimental design.

One of the specialized (and unfortunate) meanings of control group, typically used in "cross-media" experiments, refers to a group that receives some standard or "conventional" form of instruction which is compared to the treatment given to an "experimental" group that receives some new or "experimental" form of instruction, generally involving (herein) instructional media. With this specific connotation "control group" is regarded as a misnomer, and the term is not so used in the present book except when specifically referring, in quotation, to its use by others in this sense (cf. Chapter 5 , p. 27 ).[*]

We do not call this a control group, since its purpose is not to rule out an alternate hypothesis but merely to provide a comparison condition. Aside from the general deficiencies of such experiments, this group can be considered as a separate experimental treatment, even though it happens to be a conventional treatment.

A second principal meaning of "control group"--and perhaps the one the next most frequently encountered--relates to the notion of a control for "history" (i.e., concurrent extraneous influences) in assessing the effects of an instructional program; here, the control group is a *non-exposure* group for whom the treatment is "no program" or "no treatment." The purpose of such a *no exposure* (or no treatment) group--see further discussion in Chapter 9 --is to provide a "before" measure, which instead of being obtained from the *same* group of people from which the "after" measure is obtained (after being given some experimental treatment) is obtained concurrently from a separate "control" group that was *not exposed* to the treatment. This is to provide a measurement of what the experimental group *would* have been like, had they not been exposed. In this sense control versus experimental measurements are used to determine the effect achieved by a particular treatment, the *control* group being the "before" measure, the *experimental* group being the "after" measure.

A third, more generic, meaning of "control group" is in reference to any group that provides a control on any alternative hypothesis. The simplest instance is that encountered in the so-called "before-after design" (see Chapter 9 ) where one is interested in the effects of only a single treatment. The alternative hypothesis to be ruled out in this case is that the difference between "before" and "after" measurements obtained before and after treatment from a particular experimental group may have been due to some other intervening set of causes that came in between the time that the before and the after measurements were taken. Various other forms of control groups are frequently employed in more advanced scientific work to rule out other forms of alternate hypotheses. For example, when a set of practice exercises is interspersed between two sections of a film, this takes time. One might hypothesize that, rather than the exercise itself being the effective factor, merely an equivalent time used as rest period might produce the same effect; a control group which received such rest periods could be used to rule out this explanation.

It is perhaps worth reiterating, also, that a perfectly valid experiment may be conducted without a control group. The so-called control group is not a sine qua non; rather, it is something which is introduced to provide a particular kind of control that is required under the particular experimental circumstances. (One instance in which a no-treatment control group is always necessary is when we are testing a hypothesis that predicts the occurrence of differential changes in the same direction as would occur due to regression: see Chapter 15.)

A "control treatment" means a treatment administered to one or more groups for purposes of control. This term is commonly used in a specific sense when an individual subject receives more than one treatment, one of which may be termed an experimental treatment and the other a control treatment. This is frequently an alternative to using a control group and, where appropriate, may often increase the precision and efficiency of the experiment. When a subject receives both control and experimental treatments (normally these would of course have to be on different tasks or subject matters) the experimental arrangement is often described as one in which "each subject serves as his own control." The terms condition and treatment are essentially synonymous, although the term "condition" may include not only the program "treatment" but also the administration of the appropriate tests or other measuring instruments.

## Notation

In discussing experimental comparisons it is convenient to employ a shorthand notation for certain recurrent elements and relationships (see use of this in Chapter 5, pp. 40-42). The notation of experimental comparison is not well standardized (that of statistical analysis being somewhat more standard though far from fully uniform). Even within a given work such as the present book it is sometimes awkward to stick to a completely uniform meaning for a given letter or symbol throughout, and the letter X, for example, is used sometimes to represent an independent variable and at other times to stand for a dependent variable. However, within the context of discussion of experimental comparisons in which both independent and dependent variables are considered, we will generally employ the following conventions.[*]

Primary experimental variables will normally be indicated by the capital letter X. Two or more experimental treatments to be compared will be denoted by subscripts—e.g., $X_1$, $X_2$; an experimental hypothesis will normally predict that the effect of $X_1$ is greater than that of $X_2$. We will use parenthetical expressions such as "$(X_1)$" to stand for "the hypothesized effect of treatment $X_1$." Such an effect would be revealed by an observation in terms of a dependent variable, indicated by the letter, $Y_1$. The symbol "$>$" will be used with its conventional meaning, "is greater than" in relating dependent variables: thus $Y_1 > Y_2$ means $Y_1$ is greater than $Y_2$ (or higher, or better than $Y_2$). Referring to treatments, "$>$" can be read "is more effective than": thus $(X_1) > (X_2)$ indicates the *hypothesis* that treatment $X_1$ is more effective than $X_2$, whereas $X_1 > X_2$, or $Y_1 > Y_2$, indicates the *result* that treatment $X_1$ did in fact result in a superior effect in terms of the dependent measure $Y_1$.

$X_1$ can also refer to the experiemtnal *groups* of individuals that received treatment $X_1$. A "no-treatment" control (merely withholding of $X_1$) will in particular be designated by a zero subscript: $Y_o$ is the dependent measure for a no-treatment group, $X_o$.

"B/A" will refer to "Before/After," either in terms of the logic of comparison or in terms of a particular experimental pattern of procedure (See Ch. 9).

# Chapter 8
# Dependent Variables and Their Assessment

## Measurement

In experiments on educational media, we are concerned with the effects that educational media or instruments produce. More particularly, we are interested in comparing the effects on one or more dependent variables that are produced by different treatments involving one or more independent variables. The essential point is that in all instances we are concerned with a cause-effect relationship, so that the administration of a particular instructional program is of interest only insofar as its effects on the learner are observed. Further, they have to be observed in a way that enables us to determine whether we can attribute differences in effect to differences in experimental treatment. We have stressed that the ability to make such attribution is governed by the way an experiment is designed; and always included as an essential element of design is some observation of learner behavior following the experimental treatment. The present concern is to discuss some of the characteristics of this measurement, or observation, of behavior.

For our primary purpose, understanding the role of measurement in experiment, what we shall primarily mean by "measurement" is observing the effects of our treatments. An experiment must deal in observable consequents. If the hypothesized consequences are not observable, they can't be dealt with in experiments. Thus in Chapter 1 (p. 19ff.) we noted that we can deal only with behavior that can in some sense be measured (even if this means only counting of persons or instances which display some criterion of behavior that we regard as indicating an outcome sought by instruction). We also noted that such criterion behaviors may take many forms, including those we may accept as evidence of knowledge, understanding, skill, attitude, interest, motivation, etc.

Whatever kind of outcome we are concerned with, we must recognize that, although some effects of an experimental program may occur without our knowledge, we are limited in experimentation to dealing with those effects which are observed. By observation we necessarily mean here that the effect on the student is manifested in terms of some kind of overt behavior. This normally is performance on something which may be described broadly as a "test."

We should re-emphasize, however, that both "learning" and "tests" are to be understood in a very broad sense, so that the criterion behavior indicated by the "test" may, as just noted, include behavior that might often be described as indicating attitude, opinion, interest, motivation, etc., as well as changes in factual knowledge or competence in an intellectual, perceptual or motor skill. Thus in speaking of "tests" we are not necessarily talking about a conventional form of multiple-choice or other paper-and-pencil test. Examples indicating something of the range of kinds of test performance that can be assessed in instructional-media experiments can be found in the studies reviewed in Chapters 13 and 14.

One special kind of criterion to which we shall give more attention later on is the measurement, in addition to any index of competence or performance following instruction, of the time required to attain a given level of performance. This is one of the special aspects of measurement which becomes relevant only in measurement for experimental purposes, where we are interested in using measures to assess the effects of some particular treatment(s), rather than merely assessing the status achieved by a person--without reference to how it was achieved--as is more commonly the case in most educational measurement of a non-experimental sort.

Limitation of present treatment

The topic of measurement is a very extensive one, and some aspects of it are highly complex. This applies both to the theory of measurement [as an example, see Gulliksen's (1950) book on the theory of mental tests], and to the myriad details of test construction and use of tests in practical educational and psychological measurement. The considerations with which we shall deal will largely bypass both the details of measurement theory and the general technology of educational testing as such. References which may be useful in this respect are indicated in Appendix A. The experimenter on instructional media, if not already sophisticated with respect to tests and testing procedures, may have to obtain the services of a consultant on measurement as well as (as elsewhere discussed) those of a statistical consultant. Our present purpose is limited to trying to insure that the reader has in mind some essential concepts about measurement and that he is familiar with some special aspects of the application of measurement techniques that apply in experiments on instruction. This will include both some features of the measures themselves, and some aspects of the way they are used in the analysis and interpretation of experimental data.

General and special problems of measurement

Some of the problems of measurement are encountered in all scientific or technological applications in which measurement is used. Other problems are of especial relevance whenever we have to deal with highly variable data--e.g., those not having the stability of the kinds of measures familiar in classical ~~chemical~~ physics. Especially in psychological and educational measurement (but also in certain aspects of modern physics and other so-called "exact" sciences, where variability is great or only average effects can be observed), requirements for statistical analysis are posed. This applies in most experiments on instructional media, and hence in discussing experimental measurements we shall need to pursue, somewhat further than we have previously, some basic statistical concepts.

Before discussing certain problems peculiar to the analysis of changes or effects in experiments, we will first deal briefly with some general characteristics and problems encountered in all educational and psychological measurement. These are problems that arise largely because of the unreliability and indirectness of behavioral measures. As a first step, we should take a further look at the basic meaning and nature of the concept of "measurement."

The word "measurement" has many meanings. Lorge (1951), in E. F. Lindquist's "Educational Measurement" (p. 533), says:

> "Measure" is one of the thousand most common words in printed English. As is usual with words that have had a long history and wide currency, "measure" has many different meanings and applications. In a count of its occurrence in a sample of two and a half million words, "measure" occurred more than four hundred times and was used in forty different ways.

For our purposes we need not examine the forty-odd definitions. We have said that you cannot do experiments without "measurement," but we do not want to beg the implications of this statement by fastening upon some inappropriate definition. We will especially be in trouble if we fasten on the man-in-the-street's notion about scientific measurement being exact or free from error, or the man-in-the-tower's notion that measurement is a process too crude to approach the subtleties of the ephemeral qualities of the truly important aims of education.

Our concern is with doing experiments to test hypotheses about the effects of instructional media. Measurement, if required, is required for this purpose, and only in the respects that it may serve this purpose. Thus, rather than trying to proceed with our discussion of measurement from some available definition, we should examine what experiments require in the way of measurement and try to formulate a definition in terms of these requirements. Thus, our hypotheses are propositions about the effects of instructional media. Our hypotheses are assertions that under some specified conditions, certain effects will occur. Our experiments are tests to see whether the effects predicted do occur in situations designed to permit attribution to the specified conditions. Clearly, then, the experiment requires that we be able to "see," to observe whether or not the effect does occur. The observation of effect is what we call measurement. Though we will need to embellish and somewhat refine this concept of measurement, the essence of what we mean by measurement in an experiment is simply observing the occurence of behavior that may indicate effects of instructional programs. In these terms, measurement is possible if the effect is defined in such a way that its presence or absence can be recognized. It may be useful to consider that there are two parts to this requirement: (1) a definition of the effect which tells us what is being measured (2) criteria for distinguishing presence or absence of the effect.

Validity of measurement

The reason for distinguishing the two parts of the measurement requirement is that it is not always possible to work in a single stage. Sometimes it is possible to make a very good conceptual definition of the effect, but not possible to get at the defined characteristics directly. This is especially true for effects like attitudes which we might be able to define without the definition including observable behaviors. So we distinguish between direct and indirect measurement. If the definition of the effect explicitly includes observable behavior, measurement can be direct. If the definition of the effect does not explicitly include observable behavior, then we must measure indirectly through observable behaviors which reflect or are derivative of the defined effect. When measurement is indirect, the validity of the measurement is always a problem.

Whenever possible, direct measurement should be sought. The fact is that in instructional research direct measurement can be available much more often than many people think. To do so requires explicit definition of the instructional objectives. If there is good enough definition, objectives and effects become synonymous and measurement is direct.

For example, our experimental programs might be designed to teach Ss how to tie certain knots, and the effects we observe are in fact the tying of these knots as in the experiment by Roshal (see p. Ch 13), or they might be intended to teach Ss how to assemble a mechanism like an automobile distributor as in the experiments by Wulff and Kraeling (1961; see Ch. 13) or Maccoby and Sheffield (1958, 1961).* Here the behavior that reflects the desired effect of the instruction can be clearly specified in rather direct terms. On the other hand, the main purpose of an instructional program might be to develop Ss' "interest." Here the problem of specifying the kinds of behavior that will be regarded as furnishing a valid indication of the instructional objective is obviously more difficult. Some of the problems of devising and utilizing valid measures for various kinds of instructional outcomes are further discussed later.

Other qualities of measurement. In addition to the question of validity we will also need to consider some other important characteristics of measurement including reliability, discriminability, and sensitivity to change. First, however, we will consider briefly the general problem of quantification and then examine some implications of using quantified measures for the use of statistical tests by which the dependability or "significance" (i.e., interpretability as non-chance phenomenon) may be assessed.

Quantification of measurement. Let us direct our attention to the question of numbers in measurement. We do not require quantification (beyond the level of counting) in order to be able to do valid experiments, but the man-in-the-street's intuition is correct in that more exact measurements are more useful. For one thing, some hypotheses may require quantification. If the hypothesis makes a quantitative statement about effects, quantitative conditions will be dictated for the measurement. If the measurement of effect cannot meet the quantitative conditions, the hypothesis cannot be tested as stated. For example, take the hypothesis that the effect of treatment A will not only be greater than the effect of treatment B, but that the effect of A will be twice as great as that of B. We cannot test a hypothesis that specifies "twice as great" unless we have measures which are amenable to manipulations permitting statements of "twice as large." This requires not only quantification, but a certain kind of quantification.

The other and more frequently pertinent reason for quantification is that it is often required in order to achieve more precise measures. Since the error in measurement of effects is a potential source of experimental error, anything which can contribute to the precision of measurements will increase the precision of the experiment.

If the instructional-media researcher is to employ quantified measures of effect, he needs to know something of the logic underlying quantification, including the assumptions formally required for various kinds of measurement scales. These will be considered in a later section of the present chapter.

Use of quantified measures of effect

Quantification of effects starts with counting certain kinds of units of behavior. These may be of many types--for example, responses on a multiple-choice or constructed-response test of information, statements agreed or disagreed with on an opinion questionnaire, postural adjustments in a motor skill, time "on target" in a tracking task, oral reproduction of foreign language material within agreed-on limits of tolerance or acceptability as scored by judges, errors made in a procedural task, speed of performing arithmetic calculations, number of words correctly spelled from dictation, and so on ad infinitum.

Given any countable responses, we may also possibly use any of a variety of kinds of "scores" to summarize the behavior of individual subjects or of groups of subjects who were given a particular instructional treatment. The latter might include, for example, the mean or median number of correct responses made on whatever test of post-instruction competence is employed, or the percentage of Ss exceeding some criterion of satisfactory performance, or the percentage improving their scores from "pre" to "post" measures, or (with some forms of instructional programs) the time required to achieve some criterion of proficiency, or any of a number of other forms of summarization. The summarization of performance may be more compact--e.g., all performance reduced to a single value such as an arithmetic mean (which is handier to communicate but always at the expense of leaving out some of the information), or less compact--e.g., the distribution of differences in the gain scores for matched pairs of individuals or classes. Or some intermediate degree of summarization may seem appropriate--e.g., the percentages or numbers of Ss falling into each of several intervals or categories of performance, such as the percentage of "A," "B," "C," "D," and "E" grades.

Any form of summary statistics that indicate the distribution of scores (to some degree of completeness), rather than only a single summary value such as a mean, may be desired for one of two quite different purposes. The first purpose is simply that of providing a description which, while more concise than a bare tabulation of all the raw data, is more completely informative than a single summary or "central tendency" value. The other purpose is to provide information in terms of which the statistical reliability or "significance" of a summary value such as a mean may be assessed. However, not all statistical tests by any means require the use of central tendency values such as means or medians. Indeed, tests of the dependability of differences (i.e., rejection of a null hypothesis that differences in performance are reasonably attributable to chance) does not even require any quantification of individual performance at all, as was illustrated in the two examples in Chapter 4. As a first step in the interpretation of quantitative scores we may examine some hypothetical experimental results in which one sort of quantitative scores was obtained and in which the statistical test deals with the total distribution of differences between scores of pairs of subjects rather than with assessing the statistical significance of the results in terms of inferences about the stability of a mean or other average value.

## Tests of Significance with Quantified Measures

In Chapter 4 the reliability or significance of experimental findings was assessed in terms of the consistency of the findings where each outcome was simply either favorable (+) or unfavorable (-). Thus, in the hypothetical spelling-program experiment we scored as a success (+) the winning of a spelldown contest by the member of each pair of students who had the experimental program, and as a failure the opposite result that the member of the pair from the comparison treatment won out.

Before considering some of the formal qualities in problems of measurement, let us examine a similar experiment, still using the comparative performance of pairs of students (whom we may suppose are matched for ability), but where the outcome for each pair instead of being limited to a dichotomous (+ or -) is now based on their scores on a spelling test. We may suppose that one member of each of a matched pair of subjects has been assigned at random to one form of self-instructional program, $X_1$ (which, say, might be a so-called "linear"-program) and the other member of each pair has received a different form of program, $X_2$ (which, involves a form of "branching" in which words are dropped out of the program as soon as the student indicates his ability to spell them during practice exercises).

**A pilot experiment.** Let us suppose, initially, that this has been done in a small "pilot" experiment with five pairs of *Ss* assigned to each program. We will call these pairs of *Ss* A, B, C, D and E, with *Ss* $a_1$, $b_1$, $c_1$, $d_1$ and $e_1$ assigned to treatment $X_1$, and the other member of each pair ($a_2$, $b_2$, etc.) assigned to $X_2$. Let us now suppose that the scores for the five pairs of *Ss* are as follows on a twenty-word test of spelling given three days after each subject completed the experimental spelling program.

Number of words correctly spelled

$X_2$: $a_1 = 19$, $b_1 = 10$, $c_1 = 13$, $d_1 = 18$, $e_1 = 15$

$X_1$: $a_2 = 9$, $b_2 = 2$, $c_2 = 8$, $d_2 = 15$, $e_2 = 16$

Differences

$(X_2 - X_1)$: $A = 10$, $B = 8$, $C = 5$, $D = 3$, $E = -1$

Does the evidence permit us to conclude that program $X_1$ is more effective than program $X_2$---i.e., can we reject the null hypothesis that the pair differences 10, 8, 5, 3 and -1 are ascribable to chance rather than demonstrating at an acceptable level of confidence that $(X_2) > (X_1)$?

If we were to take account, as in the use of the binomial or sign test in Chapter 4, only of the direction of the differences we could not reject the null hypothesis at a level of confidence better than the .05 or even the very lenient .10 level. Thus, with one exception (the one difference with a negative sign), the chances of obtaining a difference in favor of $X_2$ as consistent as four out of five are the sum of 1/32 (the chances of getting completely consistent results with five pairs) plus 5/32 (the chances of getting just one exception), or 6/32, which is approximately .19.

The question we now face is how to take account of the further information given by the magnitude of the differences, as well as their direction. In this particular example, it is Clear that the evidence is better than would be given just by taking account of direction, since the one negative sign is the smallest value, and all of the larger values are positive in the direction that would be predicted by a hypothesis of superiority for the branching treatment $I_2$.

*Fisher's summed differences exact test.* There are, in fact, several appropriate statistical tests which might be applied. We will first consider a test which makes a minimum of assumptions about the data or the kinds of units and underlying distribution which our observed differences represent. This test, first presented by R. A. Fisher (1960, Sec. 21, p. 44) was initially described as a "test of a wider hypothesis" (as distinguished from tests which make assumptions about normality of underlying distributions) and is also sometimes referred to as Fisher's 2x2 Test, or simply as the "Fisher test."[*] In introducing this test, Fisher noted that, despite the fact that the assumption of normality is frequently not a serious limitation in practice to the use of tests into which it enters as a formal assumpation, it may "nevertheless be legitimately asked whether we should obtain a materially different result were it possible to test the wider hypothesis which merely asserts that the two series are drawn from the same population, without specifying that this is normally distributed." (1960, Sec. 21.) Fisher's discussion continues as follows: "In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied." (1960, Sec. 21, p. 45.)

The Fisher "wider hypothesis test" is based on the fact that with paired measures, as in our example, any set of obtained differences, which may have positive or negative signs and which resulted from random assignment of members of pairs to one of two treatments, would, if only chance were operating, be expected to occur with equal frequency with a positive or negative sign. That is, knowing that we have obtained differences of magnitude 10, 8, 5, 3, and 1, but with no other information, we should expect, on the null hypothesis, chance results that would yield positive and negative signs for each of these values with equal probability. By calculating all possible sums that could be obtained with either a positive or a negative sign attached to each of the five difference values, we can then ask, "What are the chances that the particular sum which was obtained in the data would be equalled or exceeded by chance?"

In our example, there are, of course, 32 such possible sums--that is, the number of sums that could be formed with either a positive or negative value for the difference between the members of each pair, or all possible sums formed by the combinations of the values +10, +8, +5, +3, +1, -10, -8, -5, -3, and -1. The arithmetical computation required to apply this test becomes exceedingly laborious if the number of pairs is very large, and even for this limited example involves 32 possible sums. However, it is evident, on inspection of the hypothetical data obtained in our pilot experiment, that we need not actually compute all such possible sums, since we can predict readily the number of these sums that would exceed that particular sum obtained for our data. This sum is of course 25, the sum of 10, 8, 5, 3, and -1. Of all the possible sums, we may ask simply how many would equal or exceed this obtained sum of 25.

Evidently, if there were no negative signs among the pairs, there would be only one possible such sum, namely 27; if there were one negative sign, there would obviously be five possible sums. But since our particular obtained negative value is the smallest of the absolute differences, it is evident that all but one of these sums would total to less than 25. On the other hand, if the one negative sign were, for example, the 3, the sum would be 21, while if any of the other differences were the one negative value, the sum would clearly be less than this. Also, it is evident that any combination whatever with more than one negative sign would also fall below the obtained value of 25. The probability of obtaining, on the null hypothesis, a value as great as or greater than the obtained one of 25 is therefore only two out of the total of 32 (this result being obtainable only if all values were positive or if the one negative value was the one which in fact it is). This probability of two out of 32, or 1/16, or approximately .06, does not quite meet the frequently adopted criterion of .05, even for the one-tailed test to indicate the chances of obtaining as extreme a result as the one obtained and in the specified direction. Nevertheless, it is materially better than that obtained from the sign or binomial test which took only direction of difference and not magnitude of difference into account.

Extending the results of a pilot comparison

It sometimes happens that the conduct of a small "pilot" experiment actually suffices to meet the objectives of the experimenter and that he can stop right there without devoting further resources to further experimentation on the hypothesis at issue. More often, however, E will use the results of the pilot experiment as a basis for deciding how to proceed further. Let us suppose that in the present case we are not satisfied with the one-tail P-value of .06 that has been achieved with one pilot experiment and thus wish to obtain better evidence by making the comparison more precise. In doing so we might seek to determine whether the betting odds favoring the hypothesis $(X_2) > (X_1)$ can be improved to, say, the .03 level for a one-tail null hypothesis (i.e., that differences as consistent as those obtained and in favor of $\underline{X_1}$, specifically, would occur by chance only four times or less in a hundred), or, on the two-tail hypothesis (that differences between $X_1$ and $X_2$ as large as obtained but without regard to which of the two was the better) would meet the .06 confidence level--i.e., would arise by chance six or fewer times in a hundred.

For one thing, before doing a further experiment we could ask whether a more refined statistical analysis might meet our purpose when applied to the data we already have. With the present data we would probably expect this not to suffice and would be unlikely to try it unless the cost of obtaining further data were very high. So we decide to get more data.

Increasing the sensitiveness of the experiment

Several methods are now available to us. If we are interested at a general level in testing the hypothesis about a method of branching we might look at the particular programs we have used and see whether in a further experiment, we might seek ways in which we could legitimately increase the sensitiveness of the comparison by increasing the efficacy of $X_2$, while still making the comparison a fair and proper test of the value of introducing the form of branching in which we are interested. Let us assume here that, however, we decide not to change $X_1$ and $X_2$, either because we are interested in the comparative effectiveness of these two particular programs as such or because considerations of cost, feasibility, etc., suggest continuing to work with these particular programs rather than seeking to increase the efficacy of the experimental variable.

We might also think of the possibility of changing the characteristics of the subject population in the conditions of utilization, with one of two objectives: either (1) to increase sensitiveness by trying to get greater differences in effect, or (2) by stabilizing the data so as to increase the consistency of the results. Here we might ask whether the effects of the experimental variable might not show up more strongly with fatigued *Ss* than with alert ones and decide to lengthen the practice sessions (if we had used more than one) or to work at a different time of day or under less favorable room conditions. Or we might feel that by using less bright subjects (or brighter subjects) the advantage of method $X_2$ might show up better (or, if we are wholly unbiased, less well!).

It is evident that were we to do so we would really be modifying our experimental hypothesis, at least in terms of a complete statement of it—though not necessarily in terms of the incomplete statement with which we started [*viz.*, "$(X_2) > (X_1)$", without reference to qualifying variables specifying the conditions under which $(X_2) > (X_1)$]. In addition, even without changing the prevailing conditions or average characteristics of the *Ss*, we might try to stabilize the comparison by procedures to obtain greater consistency of conditions, or might narrow the range of *S* to *S* variability by excluding all *Ss* deviating by more than, say, one standard deviation from the population mean in I.Q.

However, let us suppose that for whatever reasons, we decide not to try to alter either the average or the heterogeneity or "range" of characteristics of the *Ss* to be used, and that we see no readily feasible way to make the utilization conditions more suitable or more constant. (Alternatively, we might have decided to introduce some such changes but still feel that additional measures, to increase the precision of the comparison, are needed.) We therefore ask how, without changing the conditions of experimental treatment, we may increase the precision of our comparison. We have available three methods, which involve changing the experiment with respect to (1) its size, (2) its structure, and (3) the measurement of effects.

Before considering these possibilities let us note that, with the decision to keep the experiment the same except for its precision we have automatically settled for the expectation that our best current prediction about the difference in effects to be obtained is that they will be of the same order as we obtained before. We do not *know* that this will be the case (indeed we know that the average size of difference in a new experiment almost certainly will *not* be exactly the same as previously obtained). Furthermore, *if* we knew that these were the true differences there would of course be no point in doing the further experiment! Nevertheless, the single *best guess* we can make at present is to expect effects of the same magnitude, and to estimate the amount of needed additional precision accordingly. If we want to be safer, however, we can be pessimistic and guess, while still expecting an appreciable difference, that it will be somewhat less than we obtained in the pilot experiment.[*]

Improving the structure of the experiment, if we have paired our *Ss* in the optimal fashion, may not be easy.[**] Had we used a purely random assignment of *Ss* to treatments instead of pairing them in the pilot experiment we might decide at this point to pair our subjects; having done so in the pilot experiment it is less likely, though conceivable, that we would now decide to revert to an unpaired form of randomization in order to obtain more degrees of freedom [see Ch. 10, and Ch. 1, p. ]. In any case let us suppose that we decide to keep the structure the same and seek the added precision we want by increasing the size of the experiment (i.e., testing more *Ss*) or improving the discriminating power of our measures. (We might now consider doing either or both of these.)

*Increasing the size of the experiment without changing the measures.* Here, we would simply decide to repeat the experiment, either as an added replication to get more total *Ss*, including our five pilot *Ss*--e.g., five more to get a total of ten--or might think of the new experiment as replacing the pilot experiment and obtaining ten (or more) new *Ss*.* In any case, we face the question of *how many* new *Ss* to obtain. This question of predicting the number of *Ss* required to attain a given desired level of increased precision and hence confidence in the results (*if* the effects did in fact remain the same, which of course we don't know they will*†) requires a way to predict the effects of increased "N" on the precision of the comparison. This will not be calculated explicitly in the present instance, though it will be considered (in relation to a different statistical test) in a later section. For the present we might make an arbitrary decision to test 10, 15, 20, or more new *Ss*; however, first we will consider also the possibility of using a more precise or discriminating measure of effect.

*Using a more stable measure.* We might feel that the room for chance error left by our spelling test of only twenty words should be reduced. In general (see p. ) the reliability of a test will increase with the number of items in the test. By definition, a more reliable test will always reduce the amount of chance error in any given measure. (However, it will not *necessarily* reduce the amount of error in a measure of gain or difference between the scores for two conditions if some of the items, though relevant to the ability in question [here, spelling] are not differentially affected by the treatment differences.*) In addition, without changing the experimental *treatment* procedure, a difference in *procedure of measurement* might improve the precision of our measure. Thus if we used a written test in our pilot experiment we might consider using a better controlled procedure of oral testing (with control of amount of time spent per item) with less opportunity for the *S*s to use a comparison of the spelling of one word in the test with that of another, and perhaps with the use of a stop watch or other means for obtaining a measure of promptness versus hesitation, as well as of accuracy, for the spelling of each word.**

*Design of an enlarged experiment also using a better measure.* Let us suppose that, taking all of these foregoing and related considerations into account to the best of our ability, we decide to settle on an enlarged experiment with about three times as many *S*s and also a more reliable spelling test as a measure of effect. (We are not saying this would necessarily be the best decision, but merely are supposing it is the one we make.) With respect to the test let us suppose that we continue to use a written test but that we both increase the number of words in the test from twenty words to fifty words and also try to improve our scoring system. For the latter purpose we decide (whether wisely or not is a question of some interest but not relevant to our present purpose) to allow a partial-credit system in which two points are allowed for perfect spelling of each word, one point for spelling that is correct except for certain specified kinds of minor errors, and no points for incorrect spelling more serious than this.[*] We now have a spelling test with a total range of zero to 2 x 50, or 100 points (and a theoretically possible range of difference, for any pair of Ss of -100 to +100, or 200 points). Such a test will probably, though not certainly, provide a more precise discrimination of the relative spelling competence of the members of our experimental pairs of Ss, and thus help increase the precision of the experiment.

Suppose that this follow-up experiment has, in fact, been conducted, with the results given in Table 8-1 for sixteen Ss.[**] Either for convenience or as a matter of practical necessity we will suppose that these sixteen Ss came from three different fifth and sixth grade classroom groups (that being the level appropriate to our programs and--a fact which should have been specified at the outset!--also included in the range to which our hypothesis about $X_1$ versus $X_2$ is meant to apply). As indicated in Table 8-1, five of the Ss are from "5a" class, six from a "5b" class and five from a "6a" class.

In the new test of our hypothesis we shall, using the data from the paired students, work only with the differences in the scores of corresponding pair members, these having been paired up, be it noted, in advance on some relevant basis, and not paired, illegitimately, on the basis of the test data for the spelling test that serves as our criterion measure.[***] We shall thus need to deal only with the last column of figures in Table 8-1. Furthermore, we shall not have to be concerned in the statistical

## Table 8-1

## Data for a Hypothetical Enlarged Spelling-Program Experiment

| Pairs of Ss<br>No. | Class | Test Score for S given $X_1$ | Test Score for S given $X_2$ | Difference in Scores ($X_2 - X_1$) |
|---|---|---|---|---|
| 1 | 5a | 47 | 96 | 49 |
| 2 | 5a | 99 | 32 | -67 |
| 3 | 5a | 78 | 86 | 8 |
| 4 | 5a | 62 | 88 | 16 |
| 5 | 5a | 67 | 73 | 6 |
| 6 | 5b | 58 | 81 | 23 |
| 7 | 5b | 55 | 83 | 28 |
| 8 | 5b | 34 | 75 | 41 |
| 9 | 5b | 46 | 60 | 14 |
| 10 | 5b | 66 | 95 | 29 |
| 11 | 5b | 32 | 88 | 56 |
| 12 | 6a | 60 | 84 | 24 |
| 13 | 6a | 22 | 97 | 75 |
| 14 | 6a | 32 | 92 | 60 |
| 15 | 6a | 96 | 48 | -48 |
| 16 | 6a | 63 | 63 | 0 |

test of our hypothesis as stated, with the grouping of these differences by class, since, working already with paired scores we have already taken into account as a basis for controlling experimental error, and likewise as a basis for forming our estimate of that error, the correlation introduced by *all* similarities that determined pair membership, including both the relevant individual differences of *Ss* within classes and also any similar differences associated with the fact of class membership.*

Before proceeding to make a statistical test based on these sixteen pair differences we may, finally, note that the sixteenth difference, which is zero, may be considered to add nothing to the information available for testing our hypothesis.** We therefore eliminate it from the comparison leaving the following fifteen differences, which have been listed below in increasing order of absolute magnitude (i.e., magnitude without regard to sign); but underlining those values that were in fact negative:

6, 8, 14, 16, 23; 24, 28, 29, 41, *48*; 49, 56, 60, *67*, 75.

If the null hypothesis were true (that is, if the two samples of fifteen scores were random samples from the same or identical populations), these values would each be expected to occur with equal frequency with a positive or a negative sign, as was pointed out for the pilot study data. For the present data we again may ask how often a net positive sum as large as the obtained one would be obtained by chance. The maximum possible sum (which would be obtained if all fifteen values were positive) is 540. The sum we *obtained*, taking sign into account, is 314. How often, on the null hypothesis, would this sum be exceeded by chance? The total number of possible scores is now so large ($2^{15}$, or 32,768) that the labor of calculation often gets quite tedious as soon as we depart much from the simple case where all of the obtained values are positive (for which case the probability would, of course, be simply $\frac{1}{2^{15}}$ or less than .0001).* Our example has been chosen so that these fifteen values are, in fact, the same as those used in Fisher's discussion based on an experiment by Darwin (Design, Chapter 4), and Fisher's calculations for this illustrative set of differences is therefore applicable so as to permit using the results of his calculations and thus avoid the labor of doing the computation ourselves. In any case, however, one can proceed, as in our earlier simple example, without calculating all of the $2^{15}$ sums. As Fisher points out, "It is easy to see that if there were no negative signs, or only one, every possible combination would exceed 314, while if the negative signs are 7 or more, every possible combination will fall short of this value." That is, the sum of the two negative values of -67 and -48 is -115; and since this exceeds any of the single positive values, evidently with only one negative value (as well as, obviously, with none) every possible sum will be larger than 314--the largest such sum of course being if the lone negative value were 75 (which would produce a sum of 390). Similarly, starting with the smallest values, the cumulative sums of which are successively, 6, 14, 28, 44, 67, 91, and 119 for the first seven values, it is evident from the seventh value that with seven or more negative signs all possible sums would contain a negative sum larger than the obtained one of -115 (and hence that all possible sums for all fifteen values would fall short of the obtained sum of 314). Therefore we would only need to calculate the possible sums that could result when the number of negative values ranges from two through six. Even this arithmetic would be tedious; it would eventually result in the table presented by Fisher (1960, Sec. 21), the totals of which summed for zero to six negative values show that 835 such sums would exceed 314, 28 would just equal 314, and the remainder would be smaller than 314. Thus, in just 863 cases (835 + 28) out of the possible 32,768 sums there will be a positive sum as large as or larger than the obtained one of 314, the chances of this result therefore being 863 ÷ 32,768, or about .026. Since this meets the criterion we have set for the experiment, we reject the null hypothesis and conclude that $X_2$, the branching program, is more effective than $X_1$, the linear program.[*]

---

*Note: the reader may notice that the binomial test gives a lower "P" here than the present test; see further discussion in Chapter 11, p. .

Summary of illustrative experiment

Statistical test for quantitative measures. We have now illustrated in a second example the calculation of exact probabilities for a situation in which size as well as direction of differences has been taken into account. Our test has again been a "transparent" one in which all steps of the logic can be followed without special mathematical training. But the labor of the arithmetic involved (barring the feasibility of having an electronic computer and suitable computer program at hand) is so tedious for examples other than the simplest as to make it evident that a less laborious test would be most desirable [assuming we are not satisfied with the efficiency obtainable by ignoring the size of the differences and merely using the binomial table for the direction (sign) of the differences][**].

Illustrative design problems encountered. Before we look into this question we may also note that, in the course of presenting these examples we have also illustrated some of the recurrent problems of choice that arise in attempting, in designing experiments, to increase their precision, and some of the related features of design which this may involve. Thus in the particular case of trying to increase precision for a follow-up experiment after doing an initial pilot experiment we have dealt with the factor of structure, size and precision of measuring instruments, and have noted how some ways of increasing precision may also affect the validity of the measures, and/or may imply restatement of the experimental hypothesis in more complete form.

More convenient statistical tests

"There is no free lunch," it is sometimes said; you can't get something for nothing. In turning to statistical tests which are applicable to a wide range of quantitative data, yet less laborious than the exact test of Fisher just presented, we will have to pay a price.

Part of the price lies in the fact that the mathematical logic underlying such tests involves further assumptions about the data or their source which often may not in fact be fully justified. How serious or relatively unimportant this part of the price may be, for various kinds of applications, is a complex topic and has been a subject of considerable controversy among statisticians. We shall consider the matter briefly, and thus incompletely, but even that only after the nature and application of such tests has been presented.

Here we may merely note that, in the test we have just used (to determine the level of probability for rejecting the null hypothesis—i.e., the probability of getting results as consistently in the predicted direction if in fact the null hypothesis were true), we have made no assumptions about the shape of the distribution of underlying scores or their means. We have, however, implicitly assumed that the score intervals at various points along the scale from zero to 100 are equal so that the difference of sixteen between 62 and 88 can properly be regarded as larger than the difference of fourteen between 46 and 60. The latter assumption we shall have to continue to make in the further tests to be considered below, but we shall also now have to make, in principle at least, some assumptions about the nature of the distributions involved.

The other part of the price paid, in going on to these further tests, is that, for the non-mathematical experimenter at least, the logic or derivation of the test can not be fully followed--that is, the bases of such tests can not be made so fully or "transparently" clear as for the statistical tests thus far presented.

However, certain tests including the "t" test and others based on the normal probability curve are so widely used, and sometimes so convenient to use, that these disadvantages must be brooked. In the next section of this chapter we therefore try to acquaint the neophyte instructional-media experimenter with the use and structuring of the underlying logic of these tests (and perhaps also clarify some misconceptions often entertained by those with more experience). In doing so, however, the non-mathematical reader, whether experienced in experimentation or not, will have to take on faith some of the mathematical aspects of this logic which it would be far from feasible to display here as fully as was possible for the tests thus far considered.

In order to make such tests meaningful at all to the neophyte, moreover, we shall have to take a short detour through some elementary notions of discriptive statistics. The statistically less naive reader can skim through the first part of the following section, perhaps pausing here and there, however, on points which his statistics course may have left unclear (or which--as is, alas, sometimes the case--may have misled him).

## Some Statistical ABC's

As noted earlier, this book in no way attempts to serve as a text on statistical method; it is assumed the experimenter will consult his statistics text, a good statistical consultant, or both, in conducting any experiment--at least one requiring a statistical logic more complex than the simple models we have thus far described. But, at the same time, in order to understand some very basic points in the design of many experiments he may read about and some he may perform--and particularly the way experimental structure governs the valid interpretation of experimental results--the reader must also understand the basic logic and terms of reference they involve. Accordingly, in keeping with the objectives of this handbook, we hope herein to convey the minimum essentials of the logic and somewhat recurrent terms of reference that seem essential to this understanding.

We have said that a competent experimenter on instructional media, or in other fields of inquiry, need not be and probably will not be a fully competent statistician. We have, also, previously suggested why, with rare exceptions, he can *not* hope to be fully competent as statistician, and hence why he will generally need a good statistical consultant to insure validity and at the same time achieve reasonable precision for his experiments. But the experimenter must, to this end, at least be able to *talk to* the statistician (and perhaps to some degree, at least, to be able to assess the competence of the statistician with whom he is talking). Imparting even the vocabulary and statistical grammar necessary to do even the first of these efficiently goes somewhat beyond the scope of the present book; and we therefore must assume the reader will be able to refer to a basic statistics text to fill in his statistical conversing ability. All we hope to do here is to provide some "ABC's." Careful study of Chapter 4 and relevant portions of a few of the other chapters in Part II should mainly accomplish somewhat more than this (that is, more than just the ABC's) for simplest and most direct statistical arithmetical tests based on the binomial distribution and Fisher's exact probability computation. What we now need to do is to fill in some ABC's (but not much more) needed to discuss statistical tests that are based on more advanced mathematical properties.

*Samples.* The basic ABC's that we need first are a few items of vocabulary and definition. The first term we need is "sample." A full discussion of the theory of samples and sampling would easily fill several entire books the size of this one; a somewhat more extended, but still far from complete, discussion is given in Chapter 10. Perhaps our immediate purpose will be served if we merely note the following six points. (1) We never have all possible observations relevant to a given empirical question; we deal always with, in some sense, a sample of these. (2) In a truly representative sample, each member of the totality of possible relevant instances (which we will call a "population" or "universe" of instances) has *an equal chance of being included.* (3) As a special, related aspect, our notion of "comparability" of samples (see Chapter 3, p. 30ff.) means, closely enough, that any member of a population or sub-population being assigned to two treatments (the members assigned to each of which also constitute a "sample") has *an equal chance of being assigned* to one treatment or the other (and similarly for three or more samples if we are dealing with more than two treatments). (4) By "representativeness" of a sample we mean that the extent to which a sample deviates from the characteristics of the parent population from which it was drawn (or from those of a sub-group of it) is no more than would frequently occur by chance; this is properly assured only *by actually letting a chance event determine the composition of the sample* —that is, by drawing the sample by a random procedure. (5) Similarly, by comparability of two samples (e.g., those, in particular, that we wish to assign to two experimental treatments), we mean that they are randomly composed so as to differ, *de facto*, no more than would result with chance assignment. (6) The terms "population" and "sample" are not restricted to populations and samples of *people*, but may be applied to *any definable* universe of things which can be represented by a sample—thus we can speak of the mean or range of a sample of scores, of observations, of test questions, of subject matters, of spelling words, of quadratic equations, of towns, etc., etc.

*Scores.* Next, we want to consider populations of possible scores, and samples thereof (commonly, for our purpose, one or more samples representing the performance on a test given following some difference in experimental treatment) which form ordered series that are cardinal or at least ordinal scales (see p.        below). In dealing below with the "*t*-test" and (briefly) with analyses of variance, these scores will, for present purposes, be assumed to be cardinal numbers (capable of meaningful manipulation by the common arithmetical processes of addition, subtraction, multiplication, and division).

*Notation.* We will use the symbol "$\sum$" with the conventional statistical meaning of "sum of"--thus, $\sum Y$ means the sum of the Y scores. For quantitative scores in a population, as well as in a sample, we may calculate an average value--e.g., an arithmetic *mean*, M. We may use a bar over a term to indicate a *population* value, such as $\bar{M}$ for the population mean. When we use the term "average" we will ordinarily refer to the arithmetic mean. "N" will generally stand for *number* of independent observations in a sample; hence the mean of a sample, $M_Y$, of scores $Y_1$, $Y_2$, $Y_3$, equals $\frac{\sum Y}{N}$.*

## Distributions

A central descriptive notion that must be grasped in our further examination of ways we may consider the "grounds on which evidence may be reasonably disputed," to paraphrase Fisher, is that of a *distribution* or *frequency distribution*.

A distribution, or frequency distribution, refers to the absolute or relative frequency of scores with any given value. A population (or a sample) can thus be described as having $n_1$ cases of value $Y_1$, $n_2$ cases of value $Y_2$, etc., the total of $n_1 + n_2 + \ldots$ etc. in a sample equaling the total size, N, of the population or sample. Sets of scores, or distributions, are often represented graphically. The sampling binomial distribution in Figure 2 of Chapter 4 is one example. Here are two other smaller tabulations, in which each score in a sample is represented by an X in one case and, for variety, by a tally-mark (/) in the second case.

Example 1.

```
                                X
                                X        X
                       X        X        X        X
              X        X        X        X        X                 X
        Score 11       12       13       14       15       16       17       18       19
```

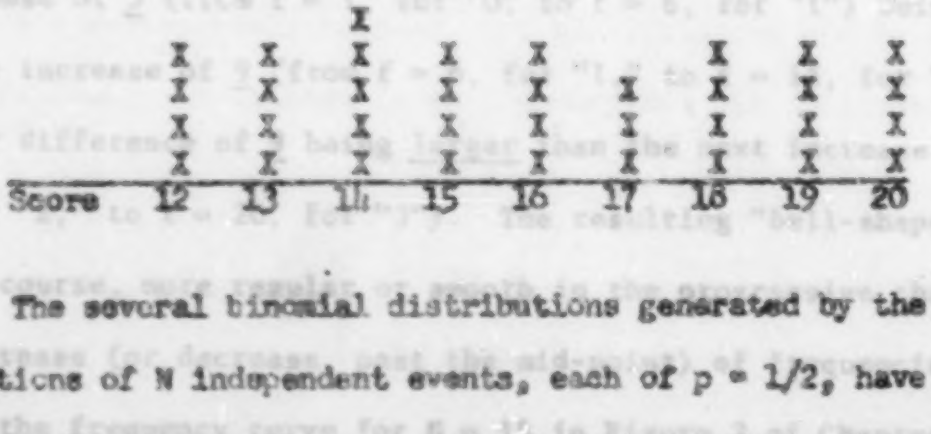
Example 2.

```
                          /
                 /    /   /    /    /
        Score  0   1   2   3   4   5   6
```

There are two kinds of distributions: obtained or <u>empirical</u> distributions (generally regarded as the province of purely "descriptive" statistics) and theoretical or inferred distributions (generally treated in relation to "inferential" statistics or scientific inference). The distributions of scores graphed above are examples of obtained distributions while the binomial distribution graphed in Figure 2 of Chapter 4 is an example of a theoretical distribution. When we speak of a "sampling distribution" of values--say, of sample means--we refer to the way means of samples are distributed. This distribution is generally arrived at theoretically, but can also be checked or arrived at empirically--namely, by getting a number of representative samples and observing how the means (or other descriptive statistical values) calculated from the samples do in fact distribute themselves.

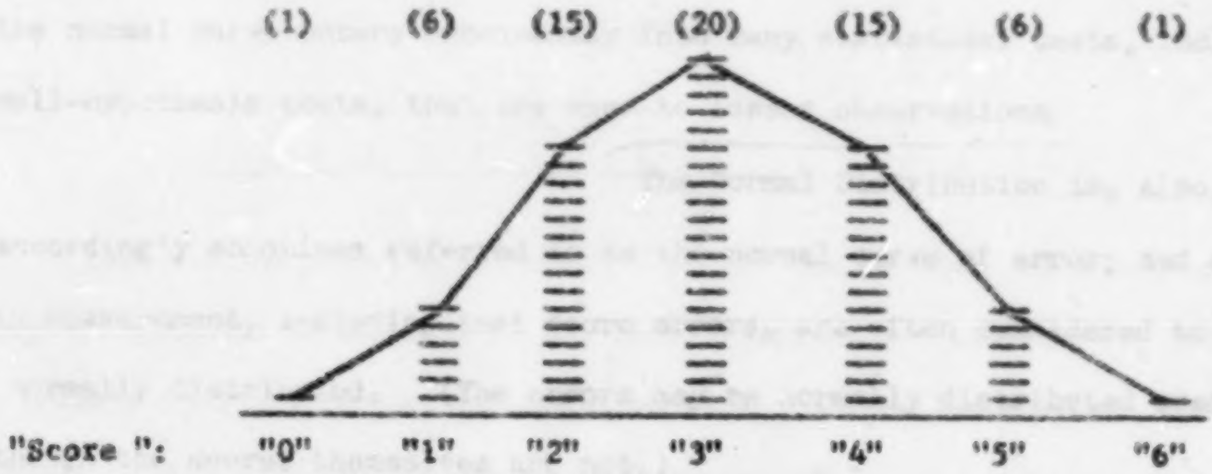
There are many important forms of distributions, only a few of which will be mentioned here, and there are various descriptive terms that refer to features of their form or nature that are of importance in certain applications but which we shall ignore here. A few key terms may be noted, however.

Certain special forms of distributions occur often enough to warrant giving them names. Distributions representing more than a few cases generally are described by the shape that the cumulated tally of cases makes. For example, the distributions on page 8-31 are roughly "triangular" in shape with a peak in the middle and "tails" at either end. On the other hand, a distribution in which each category or score value has an equal or approximately equal number of cases is called a "rectangular" distribution. The following is an example.



The several binomial distributions generated by the possible combinations of N independent events, each of p = 1/2, have been presented numerically for small up to moderately large Ns in Chapter 4.

The binomial distribution shown in Figure 2 of Chapter 4 might be described as triangular as a first approximation; however it can be more completely described. For smaller Ns the shape of binomial distributions depends on N; obviously the "50/50" distribution for N = 2 is "rectangular"; for N = 3 or 5, etc., the 1, 2, 1 or 1, 4, 6, 4, 1 frequency could be described as triangular, while for N = 2 the 1, 3, 3, 1 distribution might be thought of as a trapezoid. As N gets larger, a flexion appears (starting for N = 6) in the shape of the distribution, reflecting the fact that the rate of increase for _f_ (the frequency for each category) increases as we proceed along to the higher-valued categories 0, 1, 2, etc., and then decreases. This is illustrated below in the graphical plot for N = 6.



Here it is seen that the increase in frequency first increases, the increase of _5_ (from f = 1, for "0", to f = 6, for "1") being _smaller_ than the increase of _9_ (from f = 6, for "1," to f = 15, for "2"), but the latter difference of _9_ being _larger_ than the next increase of _5_ (from f = 15, for "2," to f = 20, for "3"). The resulting "bell-shaped" curve becomes, of course, more regular or smooth in the progressive changes in rate of increase (or decrease, past the mid-point) of frequencies, as illustrated by the frequency curve for N = 16 in Figure 2 of Chapter 4.

_The normal distribution_

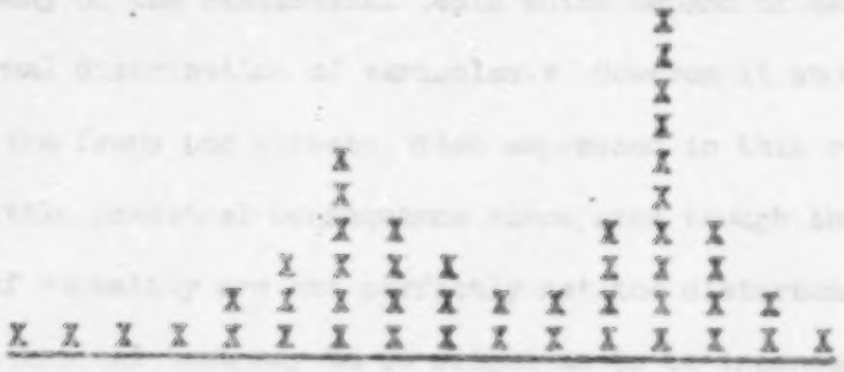
Perhaps the best known of all distributions (by name, though not always with awareness of all of its interesting traits of character) is the "bell-shaped" Normal, or Gaussian, distribution--the so-called "normal curve." Its many interesting properties can be read about in statistics texts. Here we merely note three things about it: (1) As the reader may know or might suspect from Figure 2 in Chapter 4, increasing N indefinitely in our binomial distribution would yield the Normal distribution; (2) a wide range of quantitative natural phenomena, including many test-score distributions, form approximately normal distributions; and (3) because the net effect of an indefinite number of chance events operating in different directions can result in a normal distribution, the normal curve enters prominently into many statistical tests, including null-hypothesis tests, that are used to assess observations. The Normal Distribution is, also, accordingly sometimes referred to as the normal curve of error; and errors in measurement, including test score errors, are often considered to be normally distributed. (The <u>errors</u> may be normally distributed even though the scores themselves are not.)

The normal distribution and all of the others we have considered here are <u>symmetrical</u>; however, not all distributions by any means are. The following is an example of a non-symmetrical distribution:

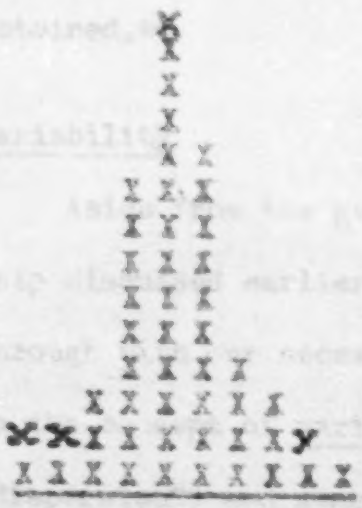
```
   X
   X
   X X
 X X X X X
 X X X X X X X X X X
```

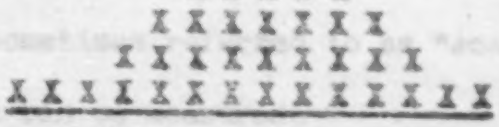
Such a distribution is said to be "skewed."

The normal and binomial distributions are also "unimodal"; that is, there is only one point at which a maximum or "mode" is reached which is higher than the other surrounding frequencies of the distribution, unlike the following distribution, which is "bi-modal" (and also, in this case, non-symmetrical):



Even a symmetrical and "normal-looking" bell-shaped distribution may not actually be normal; it may be "leptokurtic" (more peaked than a normal distribution), as in the illustration at the left, below, or "platykurtic" (flatter than a normal distribution would be), as in the illustration at the right.





A "U-shaped" distribution would evidently be the extreme case of a flat or "platykurtic" distribution.

Various statistical measures (generally based on the third or higher "moments," or powers, of the derivations of scores from the mean) are used to summarize the degrees of skewness and "kurtosis" (flatness/peakedness) of distributions. These are not relevant to our purpose here, but the reader should note that marked departures from normality in terms of skewness and, particularly, kurtosis, may—in theory at least—impugn the validity of many of the statistical tests which depend on assumptions about the normal distribution of variables.* However, it should be noted that some of the fears and caveats often expressed in this respect may in fact be of little practical consequence since, even though the theoretical assumptions of normality are not perfectly met, the disturbance to the resulting statistical test may be so slight as to be inconsequential, at least as compared to other sources or error. In a typical case the result of non-normality of the distribution might make a difference between getting a P-value (probability value) of .012 versus one of .014, which is, for the purpose of the test, a "distinction without a difference" in that the same conclusion would be drawn regardless of which value was obtained.**

### Variability

Aside from the general notion of correlation in the sense of relationship discussed earlier (see Chapter 3, p. 3-180), we are now almost through with our necessary statistical ABC's. One more essential notion is the concept of variability (also sometimes referred to as "scatter" or "dispersion") and some of the ways it can be measured.

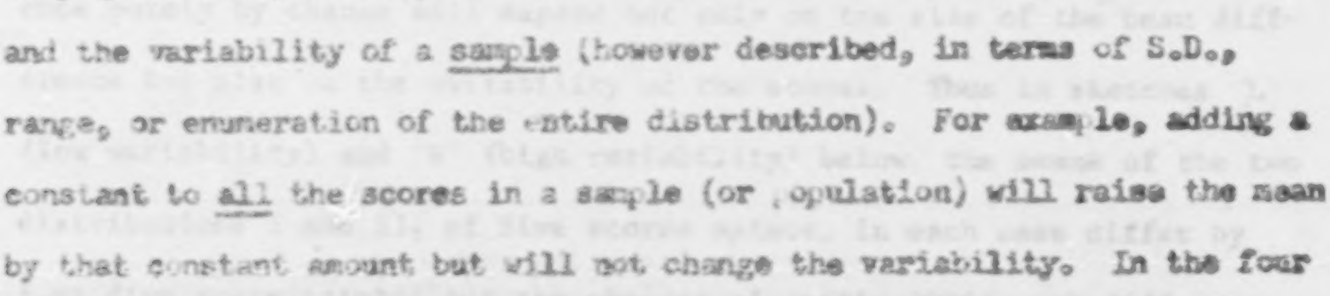
Samples of test scores or other measures—and also the populations from which the samples are drawn—are fully characterized by describing the entire distribution. When this is not convenient, we use a briefer form of description in terms of a few descriptive values. As already noted, one of these is of course the mean, a measure of central tendency of a distribution (which in a symmetrical distribution like the normal or $2^N$ binomial distributions will also be the modal or most frequent value).

The other characteristic we mainly need to use here is a measure of variability, "scatter" or dispersion. The commonest such measure is describing a distribution of scores is the range, or simply the difference between the highest and lowest score. A disadvantage of the range for describing a sample is that it is unstable--a change in just one extreme score, resulting from a sampling fluke, can alter the observed range so greatly that it is not a very satisfactory statistic in which to estimate the variability of a population based on a sample--particularly a small sample.

For reasons which we will leave to the statistics books, the generally most useful and popular descriptive measure of variability is the standard deviation, abbreviated S.D. The S.D. is also sometimes represented by the single Greek lower-case letter $\sigma$"; we shall use $\sigma$, sometimes with a subscript, to refer to particular kinds of S.D.s. The S.D. of a sample is defined as the "root mean square" of the deviations of the scores from the mean. If the scores are $Y_1$, $Y_2$, etc., we can use the lower case letters $y_1$, $y_2$, etc., to equal $Y_1 - M$, $Y_2 - M$, etc., to represent the deviations of each score from the mean. The $y$'s can be called "deviation" scores and the S.D. can then be defined as $\sqrt{\frac{\Sigma y^2}{N}}$. This "root-mean-square" definition is so basic to many statistical operations that it should be memorized. Less needing to be remembered (they can always be looked up) are various formulas with which--especially if we have a calculating machine handy--the S.D. can be computed directly from the "raw" scores ($Y_1$, $Y_2$, etc. ).* Note that, however expressed (in terms of range, S.D., or other measure), we need a minimum of two observations to form any estimate of variability. However only one observation can afford a valid estimate (even though it may often turn out to be a poor one) of the mean.

It may be noted that in a normal or approximately normal distribution the span of one S.D. on either side of the mean will always include about two-thirds (strictly, .632+) of the total "cases"; nearly all (about 95%) will be included between the points two S.D.s below and two S.D.s above the mean.* Because of these and related facts, several easy methods are available for estimating the S.D. of sizeable samples of 50 to 100 which, if the distributions are roughly normal, furnish quite a good approximation with a minimum of time and labor.**

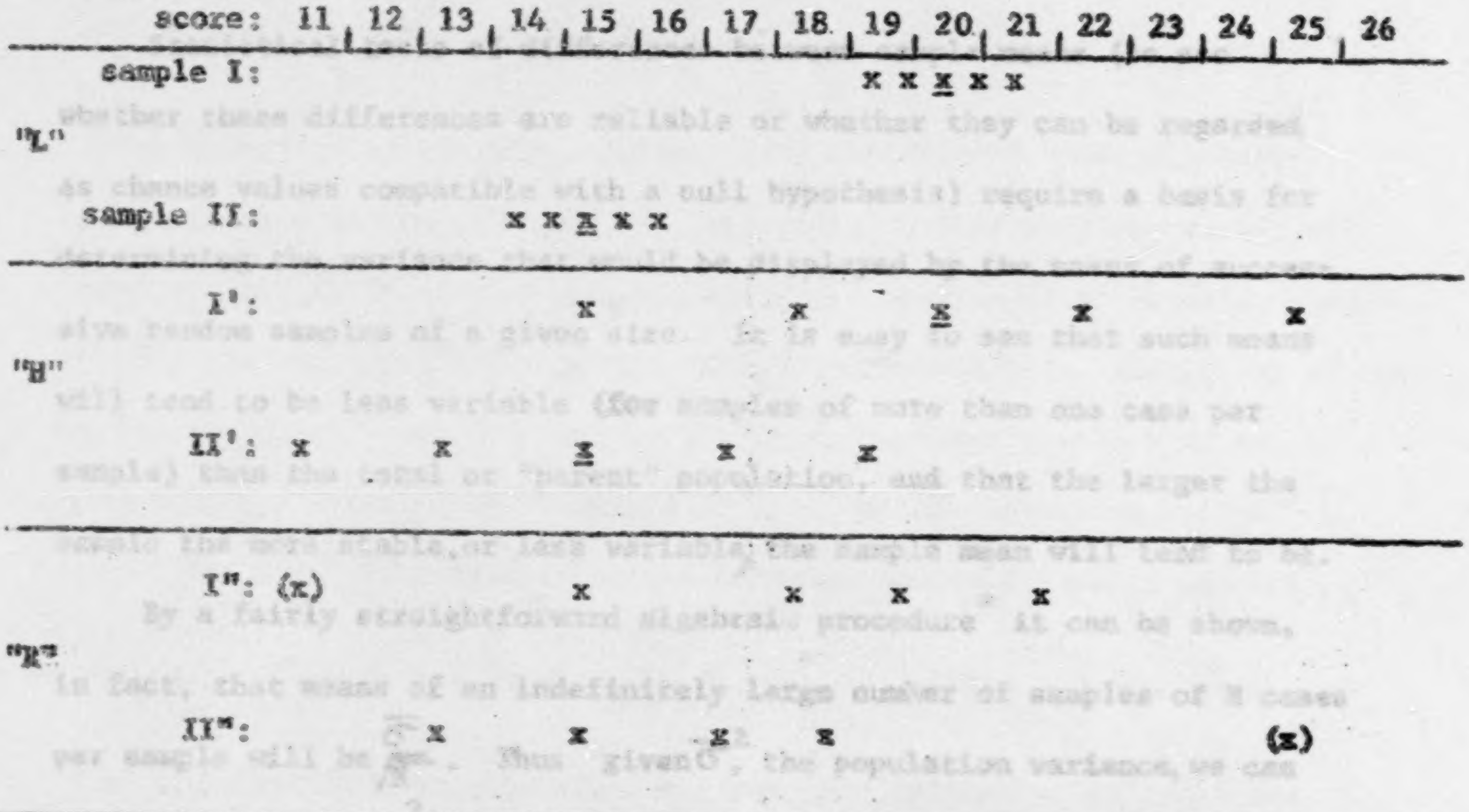
It should be noted that the variability and mean of a population may vary quite independently of each other;*** the same is true of the mean and the variability of a sample (however described, in terms of S.D., range, or enumeration of the entire distribution). For example, adding a constant to all the scores in a sample (or population) will raise the mean by that constant amount but will not change the variability. In the four illustrative pairs of distributions shown below, the two in "A" are equal in both mean (average, or central tendency, indicated by the arrows) and variability; those in "B" have the same range or variability but unequal means; those in "C" have equal means but unequal variability; and those in "D" differ with respect both to variability and with respect to mean (or other measure of central tendency).



## Statistical Tests Based on "Variance" Measures

The most commonly used statistical tests in experimental work, including instructional-media experiments, use measures of variability based on one or more S.D. ($\sigma$) or "root mean square" (R.M.S.) forms of calculation as a step in estimating the "significance" or dependability of differences due to experimental treatments. As a general basis for appreciating the general line of logic underlying the statistical tests of reliability of mean differences due to treatment effects (such as "Student's *t*-test), it may be noted that the likelihood of getting a given size mean difference purely by chance will depend not only on the size of the mean difference but also on the variability of the scores. Thus in sketches "L" (low variability) and "H" (high variability) below, the means of the two distributions I and II, of five scores apiece, in each case differ by just five score points; but the chances of getting this much difference by chance alone with as consistent values (i.e., as small variability) as in the low variability ("L") samples would obviously be slight, whereas the chances of getting the *same* mean difference with the high amount of variability in "H" would be quite high; merely interchanging one pair of scores (the 11 and the 25) in these overlapping distributions, as in the third sketch, "R", would reverse the direction of the obtained mean difference. (See the sketches on the top of page 40.)

In describing variability in terms of the S.D. or $\sigma$ as measured from a sample, or as estimated on the basis of the sample for the population as a whole, the square of the S.D. or $\sigma^2$ is often used. This measure of variability, $\sigma^2$, is called technically the *variance*. (The word "variance" is also sometimes used more loosely to simply



indicate degree of variability in a general sense, but will be reserved herein for this special measure of variability, $\sigma^2$.) The term "analysis of variance" derives from the central role played by estimates of variance in this frequently employed form of analysis. We will use $\sigma^2$ in a generic sense, for any "R.M.S." measure of variance, $\sigma_s^2$, with the subscript s (or just $\sigma^2$), to indicate the value for a sample, and the "bar" symbol, $\bar{\sigma}^2$, to indicate the population value of the variance. For a sample, $\sigma_s^2 = \frac{\sum y^2}{N_s}$, where $N_s$ is the number of observations or scores in the sample. Similarly, for a finite population, $\bar{\sigma}^2 = \frac{\sum y^2}{N_p}$, where $N_p$ is the total number of cases in the population. Since we seldom measure all the cases (there are occasional exceptions in the case of finite populations which can be enumerated or listed on a roster), we ordinarily do not know $\bar{\sigma}^2$ exactly; commonly we estimate the value of $\bar{\sigma}^2$ from one or more samples.

Sampling variance of means.

Statistical tests of differences between sample means (to see whether these differences are reliable or whether they can be regarded as chance values compatible with a null hypothesis) require a basis for determining the variance that would be displayed by the means of successive random samples of a given size. It is easy to see that such means will tend to be less variable (for samples of more than one case per sample) than the total or "parent" population, and that the larger the sample the more stable, or less variable, the sample mean will tend to be.

By a fairly straightforward algebraic procedure* it can be shown, in fact, that means of an indefinitely large number of samples of N cases per sample will be $\frac{\bar{\sigma}}{\sqrt{N}}$. Thus given $\bar{\sigma}^2$, the population variance, we can precisely predict $\sigma_m^2$, the expected sampling variance of the means of samples of any given size N, whether N be large or small. In the case of samples from populations of unknown variance, $\bar{\sigma}$ must be estimated from the sample. The relation $\bar{\sigma}_M = \frac{\bar{\sigma}}{\sqrt{N}}$ (or $\bar{\sigma}_M^2 = \frac{\bar{\sigma}^2}{N}$) is another basic relationship which the reader might well commit to memory.

Partly because of the fact that the exact treatment in the unknown-variable case was developed as a result of needs that became evident only in working with small samples, there is a common misapprehension that the relationship $\bar{\sigma}_M = \frac{\bar{\sigma}}{\sqrt{N}}$ holds exactly only for large samples, and that for small samples the denominator term should be N - 1. This is not true; the expected, or most probable, value of $\bar{\sigma}_M$ can be shown to be exactly $\frac{\bar{\sigma}}{\sqrt{N}}$ so long as $\bar{\sigma}$, the numerator, is the population value, regardless of the size of the sample. (Where the N - 1 substitution does apply is in estimating the population $\bar{\sigma}$ from the deviations of scores in a sample—as we shall see presently.)

The exactness of the relationship $\bar{\sigma}_M = \frac{\bar{\sigma}}{\sqrt{N}}$ can be illustrated by considering two limiting extremes as examples.* For the smallest possible size of sample, that is, for samples having only one case per sample, the S.D. of the "means" of such samples would obviously have to equal exactly $\bar{\sigma}$, the S.D. of the population, which in fact could be thought of as the aggregate of all such samples, each of the individual members of the population consisting of a "sample of 1" from the population. Thus $\bar{\sigma}_M$ here, for $N = 1$ (the smallest possible "small sample") equals $\frac{\bar{\sigma}}{\sqrt{N}} = \frac{\bar{\sigma}}{1} = \bar{\sigma}$; obviously it could not equal $\frac{\bar{\sigma}}{N-1}$ $\left(= \frac{\bar{\sigma}}{0}\right)$! At the other extreme, and for an infinitely large population, "samples" each of which consisted of the whole population, would always yield a mean identical with the population mean, the means of such total samples thus having no variance, as would be inferred from $\sigma_M = \frac{\sigma}{\infty} = 0$.

A further frequent misconception that should be noted is the notion that the relation $\bar{\sigma}_M = \frac{\bar{\sigma}}{\sqrt{N}}$, holds only for normally distributed populations; on the contrary, the relation is an algebraic fact that is entirely independent of the shape of the population distribution.*

Normal distribution of sample means. A further important fact is that the means of samples, distributed—as we have just seen—with a S.D. equal to $\frac{\bar{\sigma}}{\sqrt{N}}$, also tend, with samples of such size, to be normally distributed. Normality of distribution is quite rapidly approximated even for rather small samples, and this is true, moreover, for samples from populations which are themselves (the populations, that is) far from normally distributed. The means of fair sized samples from even markedly skewed distributions will distribute normally, with a very close approximation, with variance equal to $\frac{\overline{\sigma}^2}{N}$ (S.D. = $\frac{\overline{\sigma}}{N}$); the approximation is quite good enough for most practical purposes with samples even as small as N = 5 or so.* But the relationship does require that the numerator term be the population variance, or an unbiased estimate of it, not the variance of the observations in the sample. For very small samples, this requirement is of importance because the variance within the sample tends to be systematically smaller than the variance of the parent population. The discrepancy is serious for very small samples, but not very serious for large samples. We shall return to this consideration presently.

---

* Acknowledgement to F.D.S. in relation to these [illegible]

Statistical tests of means based on a known population variance. Since sample (for sizeable N's) means will predictably distribute normally with variance of $\frac{\overline{\sigma}^2}{N}$, a basis for a statistical test is afforded to see whether two given sample means differ "significantly" or reliably, or whether a given sample mean differs reliably from a known population mean, provided we know the population variance. This is seldom the case in experimental studies, including those on instructional media, but could be the case occasionally. For example, suppose the population of interest were defined as all sixth grade students in Los Angeles County, and suppose we had the data for a standard spelling test given to all of these students. Suppose further, for convenience, that this test was composed, or scored, so that the mean was exactly 100, and suppose that the population S.D. is exactly 10.0. Suppose now we give a special one-hour self-instructional program, designed to teach certain rules of spelling, to a sample of just sixteen students selected completely at random from a roster of all sixth grade students in the city, and then test them. Their mean spelling score is found to be 103.0. Assume for the present purpose that the test was well standardised, that the sixteen students' scores were unaffected by any previous taking of the test, and that conditions were controlled so that they differed systematically from students at large only with respect to having received the program. Did the program do any good? That is, regardless of whether we regard the possible gain as large enough to be important, if true, is it evidence of a true gain? That is, can we reject, at a satisfactory level of confidence, a null hypothesis which implies that the 3-point difference from 100.0 is just a chance departure from the population value of 100? (Note that this is the same question as asking if the difference between the means of 103.0 and 100.0 differs reliably from a true mean of zero.)

---

*Ref. to F.D.S., Norton, Etc.

Making the justifiable assumption that the means of random samples will be normally distributed, we can utilize the known relationship $\bar{\sigma}_M = \frac{\bar{\sigma}}{\sqrt{N}}$ or $\bar{\sigma}^2_M = \frac{\bar{\sigma}^2}{N}$, to find the expected sampling variance; here $\bar{\sigma}^2_M = \frac{(10.0)^2}{16}$, or $\bar{\sigma}_M = \frac{10.0}{\sqrt{16}} = \frac{10.0}{4} = 2.5$. We now need to assess the chances that a difference as large as the one of 3.0 we found (103.0 - 100.0) would occur just as a result of random sampling fluctuations. Obviously, it might; we might get by pure chance a few above average students in our sample, enough to raise the obtained score for the sample appreciably above the population value of 100.0. The question is, exactly, what are the chances of this occurring? We can find out by calculating the ratio of the difference, 3.0, to its "standard error"—that is, to $\sigma_M$, the expected standard deviation of sample means, which we have found to be 2.5. This ratio, which here turns out to be $\frac{3.0}{2.5}$ or 1.2, is sometimes called the "critical ratio," or C.R. It is a ratio in which the numerator is a variable, based on a sample, and the denominator is a constant, determined by the Sample size, N, and the population S.D. Since we are dealing with means which we expect to be normally distributed, this ratio can also be regarded as a "normal deviate" in a normal distribution having a S.D. of 1.00. Such normal deviates are sometimes also designated by the symbol $X/\sigma$, or by the single letter z.

The probability of obtaining by chance a "normal deviate" value of $X/\sigma$, or z, of 1.2 or larger in the expected direction can be found in tables of the normal distribution which have been worked out and which appear in most statistics books. (A partial table for values of z from 0.1 to 4.0, by 0.1 step increases, is given in Appendix B; more extensive tables, if desired, can be found in many texts.) Reference to the Normal Table shows that P, the probability of obtaining by chance a deviation as great or greater than 1.2, is 0.115*, or somewhat over one chance in ten. We are not satisfied with this as proof, and thus do not reject the null hypothesis. Had the sample mean been 105.0 instead of 103, our value of z would have been $\frac{5.0}{2.5}$ or 2.0, and the corresponding tabled "P" value would be 0.023. In this case, if the .05 level is considered acceptable, we clearly would reject the null hypothesis and, on the assumptions stated above, would conclude that the spelling program did do some good.**

A similar test could be used to determine whether two samples, exposed to two different programs, differed reliably. For example, if 32 randomly selected students had been selected, with sixteen randomly assigned to program $X_1$ and sixteen to $X_2$, we could calculate expected variance of differences between means for two equal-sized samples of the size used. This (see also Chapter 10, p. 10- ) can be shown to be $\frac{2\,\overline{\sigma}^2}{N}$ or, for our particular $\overline{\sigma}$ of 10.0, $\sigma^2_{\text{mean diff.}} = \frac{2(10.0)^2}{16}$, or $\sigma_{\text{M.D.}} = \sqrt{\frac{200}{16}} = \frac{14.1}{4} = 3.53$. Mean differences* between the paired scores of students in samples that were given the two treatments could then, as can be seen by reference to the Normal-distribution table, meet the criterion of P as low as .05 for obtaining a difference in a specified or predicted direction if the difference in means were about 5.6 or greater, which would yield a value of "Z" or "$x/\sigma$" of $5.6 \div 3.5$ or 1.6, for which, from the table, P (one tail) is found to be .045. [If, however, we only asked which program is better (without specifying direction of expected superiority) and were thus prepared to accept a difference in either direction as evidence of superiority for one of the programs, we would then, still sticking to a "confidence level" of .05, have to require an $x/\sigma$ or $z$ value of nearly 2.0--for which P for one tail is seen in the normal table to be .023, so that the "two-tail" P, for a difference in either direction is about .046.] (For the more stringent .01 level we should need an $x/\sigma$ value of about 2.3 for the one-tail or unidirectional hypothesis, and about 2.6 for the bi-directional or two-tailed test. (See Table, p.534).

The need for statistical tests based on parameters estimated from sample variability. In using the binomial or sign test, the only value to be calculated is the number of discrepancies or failures;

the rest of the statistical test follows from a theoretical distribution that has been derived or can be derived independently of the data, given only the size of the sample, N. The same applies in testing the reliability of proportions or percentages (or percentage differences) of various sorts--given the sample "mean"--i.e., the obtained percentage, its statistical reliability or "significance" can be obtained from a known theoretical distribution.*

Somewhat similarly, in determining the probability that a given sample mean difference deviates significantly from the population mean, or that a difference between samples deviates significantly from a true mean of zero, we need only calculate the mean (or difference in means) from the sample(s) and test it against a predictable theoretical distribution for means of samples of that size, provided --as in the examples above--the population variance is known.

However, there are two important kinds of situations in which this type of calculation, of comparing a single observed value directly with a known distribution, can not be performed, and--unfortunately--these kinds of situations are among the most frequent with which the instructional-media experimenter will have to deal. The first kind of situation is where the form of theoretical sampling distribution is known, all right, but one of the population parameters that is needed to employ the distribution is not known a priori (even though it can perhaps be estimated from the data themselves). This situation occurs in the use of the well known t and F tests and is discussed below. The other situation is where the appropriate theoretical distribution is simply not known--i.e., has not been worked out mathematically. This may be either just because it just hasn't been, because it is too difficult for available mathematical techniques, or because the distribution involves inconvenient or awkward properties so that it would not be serviceable even if worked out. An example of this last sort is found in the case of the "effectiveness index"* discussed in a later section.

Comparison of means when the population variance is unknown

Perhaps the commonest situation encountered by the instructional-media experimenter is that where he wishes to compare the means for two samples (or to test the dependability or "significance" of a mean difference between two sets of paired scores) where, as is almost always the case, the population variance is not known, and where, therefore, a test using ~~in~~ "$D/\sigma$" (of ratio form) will require that E must estimate $\sigma$ from the variability of his sample or samples. Now N, the size of the sample, will not only affect directly the expected variance of the mean, but also will affect the estimate, $\tilde{\sigma}^2$, of population variance which he must use for such a test. In this case the estimate must take into account not merely the size of the sample as such, but the number of independent observations, or "degrees of freedom" on which the estimated population parameter, $\tilde{\sigma}^2$, is based.

"Degrees of freedom"

In the statistical or probability tests discussed in connection with the hypothetical experiments in Chapter 4, we pointed out that the validity of the probability calculations that composed the test depended on reckoning probabilities for independent events. These and other statistical tests are set up to possibly disprove a null hypothesis which predicts the consistency or amount of difference that would be likely to result just from chance. The number of events (e.g., pairs) independently assigned to two treatments governed the selection of the appropriate frequency distribution. In the kind of experiment dealt with in the second experiment in Chapter 4 it was the number of pairs, each of which independently had one of its members assigned to each treatment, that determined which binomial distribution--that is, which power of two--was appropriate. Thus, for, say, eight pairs we used the $2^8$ distribution; for ten pairs the $2^{10}$ distribution, etc.

The number of independent acts of randomization in assigning pair members to treatments in this case can be called the number of degrees of freedom in the statistical test. The term "degrees of freedom," often abbreviated d.f. or d/f, will be frequently encountered also in other statistical tests based on a more involved mathematical derivation than the one we spelled out in the second part of Chapter 4 for the sign or binomial test.

In the present discussion we will use "d.f." to stand for the words "degrees of freedom" and "d/f" to mean "number of degrees of freedom." In any statistical test the d/f (number of degrees of freedom) will determine the appropriate theoretical frequency distribution which will predict how observations would, on the average, distribute. In some tests, also, the number of degrees of freedom enters into the test as a denominator term in obtaining a ratio, estimated from the data, which will be compared to the theoretical probability distribution. This is the case, in particular, with the well known t-test, based on the ratio known as "Student's t" (see p. 53 ) and with the "F" ratio used in the various forms of the analysis of variance.

In all cases, regardless of the form of the test, the number of degrees of freedom (d/f) in a valid test is (as was the case in our binomial test in Chapter 4) generated directly by the number of subjects, of pairs of subjects, or of intact groups of subjects, that are independently assigned to one or another experimental treatment. However, the d/f (or number of d.f.) does not always exactly equal the number of Ss (or pairs, or groups, of Ss) randomized; it may, depending on the test to be used, be slightly smaller than that number. In certain instances, particularly when the d/f is small, this difference in size may be of critical importance.

The further tests we shall deal with will, as in the test just considered, mostly involve computations of averages in the form of arithmetic mean (i.e., the sum of a set of scores, and with the sign "$\Sigma$" standing for "summation of," the mean, M, of N scores = $\frac{\Sigma Y}{N}$).* The t-test is a test for determining whether two means, $M_1$ and $M_2$, differ beyond chance expectancy, or whether a given mean, M, differs significantly from zero, when the population variance is not known. For the latter of these situations (Is M $\neq$ 0?), the d/f = N - 1. For the former (Is $M_2 > M_1$, or is $M_2 - M_1 \neq 0$?), where the two means are based, respectively, on $N_1$ and $N_2$ independent observations, the d/f = $N_1 + N_2 - 2$; see also Chapter 10, p. 10- . (The number of d.f. in other statistical tests can be found in appropriate parts of textbooks in statistics describing these tests; in Chapter 11 we will consider, but only briefly, one or two of these further tests that involve the "analysis of variance.")

The reason that the d/f involves the subtraction N - 1 in a test involving an $x/\sigma$ ratio, where $\sigma$ has to be estimated from the data instead of being known a priori, is that a $\sigma$ based on N observations, being computed from deviations from a sample mean, no longer is based on N independent observations. For example, consider the sample of five values such as the five pairs of differences in our pilot experiment (see p. 7-12). These values were 10, 8, 5, 3, and -1. Their algebraic sum is 25 and their arithmetic mean is 25 ÷ 5 or 5.0. Now, in calculating a measure of the dispersion or variability of these values in terms of the squared deviations from their mean (which we wish to do in order to form an estimate of the true variance of population from which these particular values are a sample), we do not have five independent deviations but only four--because, once we have determined any four of the individual values and their mean, the remaining value is automatically determined. (That is, if we know the mean of five values is 5.0, the sum must be 25; and given that four of the values are, say, 10, 8, 5, and -1, which total to 22, the remaining value can only be three.) One degree of freedom for calculation of variability has thus been lost in basing the variability computation on the sample mean, and we have therefore only four independent deviations, or d.f., for the variance calculated from our five cases. More generally, there will evidently be N - 1 d.f. available for estimating variance from a sample of N independent observations.

## Unbiased estimate of population variance

As a corollary of this fact, the best estimate of the population variance $\tilde{\sigma}^2$, from the observations in a sample of size N, is not the same as the observed variance of the sample. It was once thought to be so, and little harm was done through this error as long as statisticians were concerned only with large samples. One of the important contributions of W. S. Gosset ("Student") came about by his finding, in effect (through study with repeated samples obtained in his work in a brewery) that the $\sigma_s^2$ or S.D.'s for samples tended to be smaller than the population values, and that estimates of the latter, based on calculation of the actual variance found in samples, would thus be biased in the direction of underestimating the true variance.

An unbiased estimate, $\tilde{\sigma}^2$, of a population variance is, when based on a sample of size N, exactly $\frac{\Sigma y^2}{N-1}$ (rather than exactly $\frac{\Sigma y^2}{N}$). Contrary to what is sometimes believed (and even taught) this applies to large samples just as much as to small ones; however, the discrepancy will obviously be a slight one, for samples of any considerable size.*

We will use $\tilde{\sigma}^2$ and $\tilde{\sigma}$ (with a wavy line) to indicate an estimated population variance (= S.D.) based on a sample. [Note also that the subscript M in $\sigma_M^2$ (including the more precisely designating forms $\tilde{\sigma}_M^2$ and $\bar{\bar{\sigma}}_M^2$) indicates not the variance of individual scores but rather the sampling variance of the means of samples.]

Of even more importance in the case of small samples ** than the use of the proper denominator N - 1, equal to the true d/f, to yield an unbiased estimate of the population variance, is another fact which also applies logically to large samples, but which, like the discrepancy between the N and the N - 1 as the denominator term in $\sigma^2 = \frac{\Sigma y^2}{N-1}$, also is not conspicuous with largish samples of more than about 20 or 30 observations. To appreciate this second fact we may recall that sample means, and therefore the fraction $\frac{M}{\sigma_M}$, tend to be normally distributed when $\sigma_M$ is based on the known population variance, $\bar{\bar{\sigma}}^2$.

Since $\bar{\bar{\sigma}}^2$ is then a constant, $\frac{M}{\sigma_M}$ must likewise be normally distributed, in this case, for any size sample, if M is thus distributed, ~~the variability of the sample accordingly used as basis~~. But in the much more usual case where $\bar{\sigma}^2$ is not known (and where for any test based on an "$\frac{M}{\sigma}$" form of statistic must therefore use an estimate, $\tilde{\sigma}^2$, based on the variability of sample observations), the denominator of such a ratio is a variable as well as the numerator, and the ratio is not normally distributed (even if the parent population is)—even though for quite large samples normality will be well approximated.

*The t distribution.* Gosset's contribution, which put scientific inference for the first time on a rigorous mathematical basis, was to discover the true sampling distribution of the ratio $\frac{\tilde{M}}{\tilde{\sigma}_M}$, where both $\tilde{M}$ and $\tilde{\sigma}_M$ are unbiased estimates of the population values of $\bar{M}$ and $\bar{\sigma}_M$ based on samples of any given size. The sampling distribution of this ratio, designated as "Student's t," or simply t, changes with the size of the sample (or, more directly, with the d/f on which the estimate $\tilde{\sigma}^2$ is based). The curves in Figure 8-1 show the distribution of t for d/f of 1, 3, 10, and 20 in terms of the probability of obtaining by chance a value of t as large as, or larger than, a given value, for a difference in a specified direction. (These curves are plotted in Fig. 8-1 with the P-values as ordinates on a logarithmic scale in order to provide adequate differentiation for the lower P-values.) The left-hand curve, labeled "$\infty$," indicates the values for the normal distribution, which is exactly the same as t when the d/f become infinitely large and, as seen, is very closely approximated for d/f over 20. Comparison of the curves for 1, 2, 3 or 4 d/f with those for the higher d/f of 20, 10 or 7 makes it apparent that the discrepancy from the normal distribution is very great for very small samples, but that the difference rapidly gets less as soon as the d/f reach a moderate number.

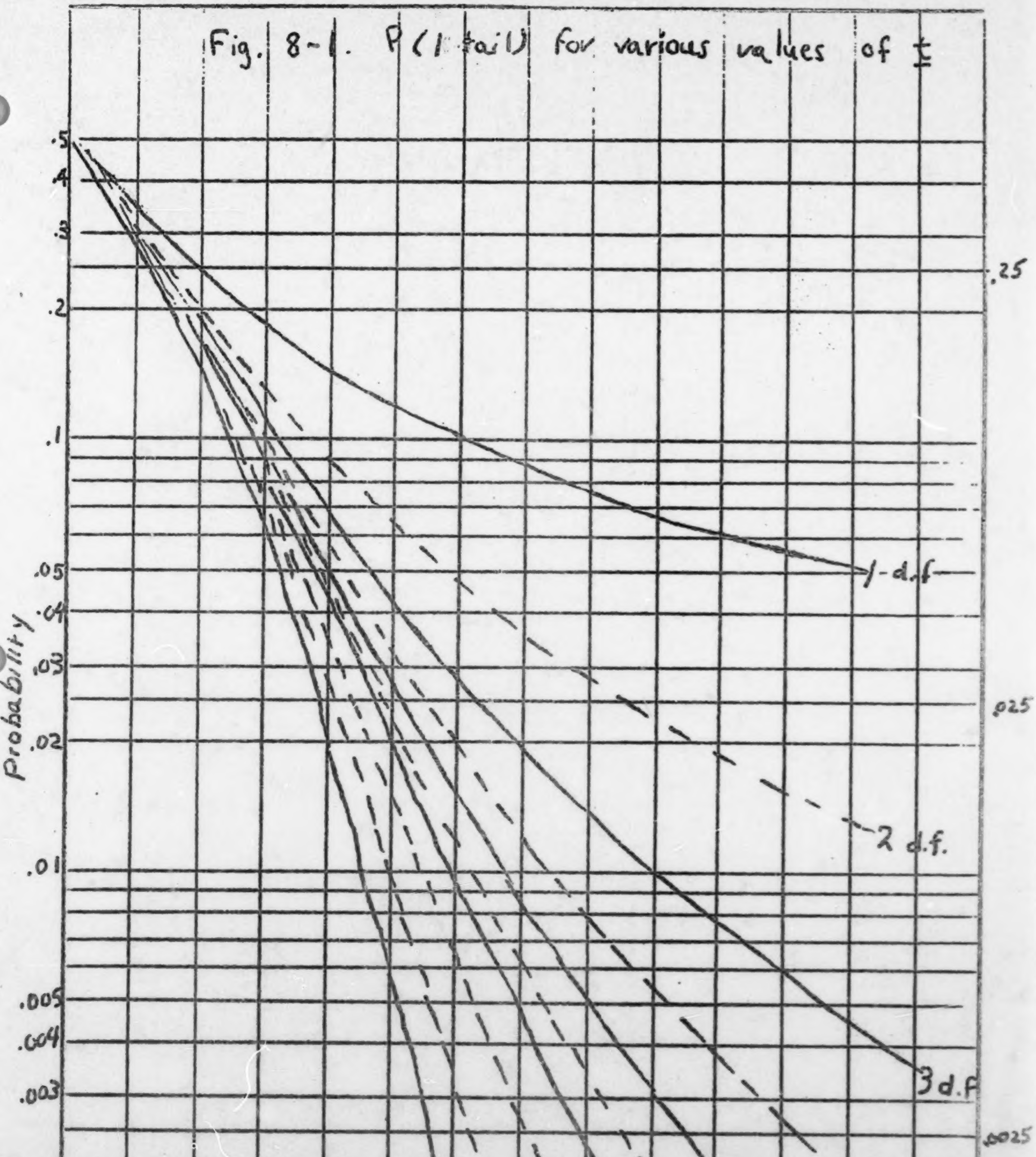


Fig. 8-1. P (1 tail) for various values of t

Note also that, at least for more than 1 or d/f, the discrepancy between the normal and the t distributions is less serious for the more "lenient" P-values (e.g., P of .05, one tail) than for the more stringent rejection levels (e.g., .005 and below). Thus for a moderate value of $t = \frac{x}{\sigma_M}$ of 2.0, we have P = .07 for 3 d/f and P = .05 for 5 d/f, as against a P of only slightly under .025 for the normal curve; thus, given a t based on an unbiased estimate of $\sigma^2$, the conclusion reached by using the normal distribution instead of the proper t distribution would not be seriously wrong in such a case.[*] However, at the other extreme of very small d/f and higher t (and thus generally lower P-levels), we would obtain, for a t of, say, 4.0, a P-value of <.001 for the normal curve (as well as for the 20 d/f t curve), and thus surely reject the null hypothesis, whereas for only 1 d/f we would, with the same t, obtain a P of nearly .08 (.15+ for two tails) and thus very likely would not reject the null hypothesis.

The underlined values in the following supplementary table are useful for interpreting the higher values of t for very low d/f of 1 and 2 which could not be shown on Figure 8-1 without extending its size (or compressing its scale) unduly. (Here the symbol P' stands for the one-tail value and the symbol P" stands for the corresponding two-tail value.)

| P' | .05 | .025 | .01 | .005 |
|---|---|---|---|---|
| P" | .10 | .05 | .02 | .01 |
| 1 d/f | 6.31 | 12.71 | 31.82 | 63.66 |
| 2 d/f | 2.92 | 4.30 | 6.96 | 9.93 |

A more complete tabulation of t is given in Appendix B.

The reader should note that the distributions given here, like those originally published by "Student",[**] are for P' (one tail of the distribution), and the table gives these P' values for specified values of t spaced at even intervals, whereas the tables in most textbooks differ in that, following Fisher's practice, they instead (a) employ the P-values (P") for *both* tails of the distribution, and (b) use the converse arrangement of giving in the body of the table the required value of $t$ for set values of P (usually including P" = .05 and .01). Confusion can ensue for the unwary reader, particularly since many writers of research reports, some of whom are unaware of the difference, report P-values without specifying which P-value (one tail or two tails) they are reporting--and often they omit giving the values of $t$ (or other statistics) from which this could be determined.[*]

*The one-tail, two-tail wrangle*. In Chapter 4 (p. 59) it was noted that considerable discussion, and--in the writer's opinion--much unnecessary confusion, has centered around the appropriateness of using the one-tailed (or one-sided) test versus using the two-tailed (or two-sided) test.[**] Actually the "test" is in a sense the same; it is primarily a question of what P-value to report.[***] Positions adopted vary all the way from advocacy of invariably reporting the two-tail value (which, as noted above, is the one most commonly tabled in statistics texts), on the grounds that this is "more conservative," to pronouncements that one should use one tail if he "has a hypothesis" that predicts the expected direction of a difference (sometimes abbreviated to merely "if he has a hypothesis") but two tails if there "is no hypothesis." We shall not enter here into an extended discussion of the arguments that have been advanced for various positions on the one-tail, two-tail controversy. These points may be made, however.

1. If the relevant statistics (e.g., $t$ and d/f) are given--as they should be--the reader has the choice, if he prefers and knows what he is about, of doubling or halving, as the case may be, the reported value to suit his fancy.[****] To make things slightly easier, it would be handy if *Es*

would make it a practice to label their reported values as one-tail or two-tail values, possibly adopting our convention of using the symbol P' for the one-tail value and P" for the two-tail value,[*] or of using such expressions as $P < .01$ (.02) to indicate both of the two values.

2. In terms of logical nicety, more than (usually) of direct practical consequence, we would contend that the one-tail test is generally appropriate when, as we shall further argue should generally be the case in instructional-media experiments (cf. Chapter 12), *E* has a theoretical basis for expecting $(X_1) > (X_2)$, rather than merely asking the "fishing-expedition" question, "Do $(X_1)$ and $(X_2)$ differ?".[**]

3. As noted in Chapter 4 (p. 59), the difference of a factor of two in the resulting P-value is, as Mosteller and Bush (1954) suggest, "scarcely worth arguing about." The interpretation to be made generally hinges more nearly on the order of magnitude of the P-level than on its exact value.[***]

## The logic and assumptions of the "*t*" test

The logic of the *t*-test is essentially that of testing a null hypothesis to see either whether (1) two sample means $\bar{Y}_1$ and $\bar{Y}_2$ differ "significantly," or reliably--i.e., more than we are willing to ascribe to chance; or (2) whether a given sample mean, $\bar{Y}$, differs reliably from zero. (As just seen, the single mean can be the mean of the differences of two sets of paired observations.) In either case, certain assumptions are made about the scores on which the mean or means ($\bar{Y}$, or $\bar{Y}_1$ and $\bar{Y}_2$) are based. If these assumptions are not fully met, the test is, strictly speaking, inapplicable; however, experience shows that in practice some of the theoretical assumptions may be departed from rather widely in certain respects without its making much practical difference in the outcome (cf. also Chapter 11, p. ).

The reader's statistical consultant should be the judge of how unsafe the use of the _t_-test (or normal-deviate test, if population variance is known) may be for a given kind of data obtained in an instructional experiment; but we would emphasize that the _t_-test actually has very considerable latitude and can be safely used for a variety of scores that would not, in a pure sense, meet the assumptions made in the mathematical derivation of the _t_-test sampling distribution. Thus, despite what is sometimes believed, the _t_-test does work quite well, given a roughly equal-interval scale, even though the population distribution may be quite far from normally distributed (purist cautions to the contrary notwithstanding).*

Some discussions of assumptions underlying the _t_-test stress the elements of normality, particularly with respect to skewness, though Norton (1952) has shown that marked departures from normality (in the parent population distribution), with respect to kurtosis, tend to be the serious cause for concern.** However, perhaps the most important assumption (often not mentioned) is that the scores approximate a scalar variable with equal (hence, properly additive and subtractive) units. If this assumption is grossly violated--that is, if in some sense the difference between two scores at one part of the scale is not rougly comparable to an arithmetically equal difference elsewhere in the scale--then any statistics based on sums and means, including those used in the _t_-test, are suspect. (The same applies, furthermore, even to most of the so-called non-parametric tests; see also Chapter 11.)

The logic of the _t_-test depends on the theoretical sampling distribution for a ratio, $\frac{\bar{Y}}{\sigma_{\bar{Y}}}$ , or $\frac{M}{\sigma_M}$--that is, a ratio of an observed mean, based on a sample, to an estimate, also based on a sample, of the variance to be expected for means of samples of that size. The logic is the same if we are concerned with a difference between the means $M_1$ and $M_2$ of two independent samples (see Chapter 10, p. 22 ff.); for simplicity in illustrating its application, we will stick to the case of a single mean, M, and the null hypothesis test that it can be regarded as a sample from a population whose actual mean is zero.[*] This is the case, of course, when we have two sets of paired differences and ask whether the mean difference differs reliably from zero. As an illustration let us apply the test first to our 5-S pilot experiment and then to the data from the follow-up experiment.

### Illustration of t-test for paired differences

The computation of t is simple in principle, and the computation itself is not very laborious, especially when we are dealing with a small number of observations. As an example, we can first take the data from our "pilot" experiment as presented on p. 8-12, where we obtained the differences 10, 8, 5, 3, and -1 between the scores of our five pairs of Ss. The algebraic sum of these differences, which we may call $\Sigma D_{Y_1 - Y_2}$ (or simply $\Sigma Y$), is 25.0, and their mean is thus 5.0. The following tabulation shows the deviations of each difference from the mean of 5.0, the squares of these deviations, and the sum of these squares:

| Pair | Diff. in scores (Y) | $y = Y - M_Y$ | $y^2$ |
|---|---|---|---|
| A | 10 | 5 | 25 |
| B | 8 | 3 | 9 |
| C | 5 | 0 | 0 |
| D | 3 | -2 | 4 |
| E | -1 | -6 | 36 |
| Sum | 25.0 | 0.0 | 74 |

Note that the algebraic sum of the deviations from the mean is zero, as it must be, by definition since $\Sigma y = \Sigma(Y - M_Y) = \Sigma Y - \Sigma M_Y$, and "$\Sigma M_Y$" is simply N times $M_Y$, or $N \frac{\sum Y}{N} = \sum Y$); this fact obviously provides a useful arithmetical check.

Now, in order to determine by means of the t-test whether we can reject the null hypothesis that the mean of 5.0 is merely a chance deviation from zero, we can compute t, in four steps.*

First, we obtain an unbiased estimate of the population variance, $\tilde{\sigma}^2$, using the relation $\tilde{\sigma}^2 = \frac{\sum y^2}{N-1} = \frac{74}{4} = 18.5$.

Next we use this value of $\tilde{\sigma}^2$ to obtain the expected variance of sample means (for samples of N = 5), that is: $\sigma_M^2 = \frac{\tilde{\sigma}^2}{N} = \frac{18.5}{5} = 3.70$

Third, we divide our obtained mean difference by its "standard error," $\sigma_M$. Thus: $t = \frac{M}{\sigma_M} = \frac{5.0}{\sqrt{3.70}} = \frac{5.0}{1.92} = 2.6$.

Finally, we look up the probability of obtaining a t this large, using the values for N-1 d/f , in this case 4. Reference to Fig. 8-1 (or to a table of the distribution of t) shows that the "one-tail" probability for t = 2.6 with 4 d/f is about $P' = .03$, which is the probability, given the assumptions of population normality underlying the t distribution, of obtaining a mean difference as large as 5.0, and in the specified (positive) direction.** The reader might note that this value of .03 is somewhat better than the one of .06 obtained for this sample of data with Fisher's exact test (p. 8-15) for obtaining a difference-sum as large as or larger than the obtained one of 25.0; it is, in fact, about the same as the previous calculations would have yielded for getting a difference-sum larger than the obtained one (which would be 1 in 32, or .03).***

For the larger set of observations in the follow-up experiment, the agreement between the two tests is closer. If, as before, we ignore the one

---

*** Yates correction.

zero difference in Table 8-1,* the mean of the 15 differences (see p. 8-22) is 20.93. For the calculation of *t*, with this many observations and a less convenient mean than in our pilot data (so that the deviations from the mean are, as is generally the case, unfortunately, not integers), we would find it handy to use a calculating machine, a table of squares and square roots, and perhaps a computational formula adapted for efficiency in machine calculation. Direct computation, in terms of the basic formula that we are using here for simplicity, would give:

$$t = \frac{M_D}{\sqrt{\frac{\Sigma y^2}{N \cdot N-1}}} = \frac{20.93}{\sqrt{\frac{19,945}{15 \cdot 14}}} = \frac{20.93}{\sqrt{94.97}} = \frac{20.93}{9.75} = 2.15,$$

for which, with 14 d/f, P' (P for one tail) can be found, by interpolating between the curves for 10 d/f and 20 d/f in Fig. 8-1, to be about .025, agreeing very closely with the value of .026 found by the Fisher exact-probability test (see p. 8-41).**

## Analysis of variance

Basically, the "analysis of variance" can be thought of as an extension of the logic of the *t*-test. Some might argue with this characterization, but we believe it is the point of view best calculated to promote clarity at least for the neophyte.***

Since we are not trying to provide a statistics text, we will confine our attention to a few simple purposes and forms of the analysis of variance. Of these, some in fact can be seen to reduce directly to the *t*-test; others essentially extend its logic to either (1) simultaneous comparison of the effects of more than two treatments--say $X_1$, $X_2$, and $X_3$ of a given variable, X; or (2) simultaneous analysis of the effects of two (or more) variables, one of which might have been used as a "sorting" or "stratifying" variable (see Chapter 7, p. 28). Most of our attention will be devoted to the latter of these cases; the former we attempt to characterize briefly below.

Suppose that we have three (or more) categories of treatment, say $X_1$, $X_2$, and $X_3$--say, degree of prompting or amount of repetition in an instructional program. The three categories need not form a quantitative series, however; they might represent qualitative, non-ordered categories such as *kind* of cueing or *kind* of repetition. We want, instead of asking merely, "Is $(X_1) > (X_2)$?" and going on to inquire separately, "Is $(X_2) > (X_3)$?" and "Is $(X_1) > (X_3)$?" to determine whether factor X, as represented by these three conditions with respect to it, makes a difference. That is, we want to test the null hypothesis that variation in one dependent measure among the three treatment categories is not greater than zero except within the range readily ascribable to chance.

Taking only the simplest instance where we have N independent observations apiece for $X_1$, $X_2$, and $X_3$ (e.g., N *S*s apiece randomly assigned to each treatment), as analysis of variance to test this null hypothesis is, essentially, a comparison of two ways of estimating a population variance: (1) from the variation among scores of *S*s *within* treatments, and (2) from the variation exhibited *between* treatments--that is, among the three treatment means (call them $\bar{Y}_1$, $\bar{Y}_2$, and $\bar{Y}_3$).

The logic for (2) is essentially the inverse of that described above for predicting the variability of sample means, That is, given $\tilde{\sigma}^2$ (or $\bar{\sigma}^2$), we can predict $\tilde{\sigma}_M$ by the relation $\tilde{\sigma}_M = \frac{\tilde{\sigma}}{\sqrt{N}}$ (cf. p. 41, above). So, conversely, from the *observed variation* of sample means (that is, of $\bar{Y}_1$, $\bar{Y}_2$ and $\bar{Y}_3$), we can work backwards and estimate $\bar{\sigma}$, the population S.D. Actually, for convenience we use the $\sigma^2$'s or variances, so that, since the variance of sample means is then given by $\sigma_M^2 = \frac{\bar{\sigma}^2}{N}$, the converse relation is $\bar{\sigma}^2 = N\sigma_M^2$.

The logic for (1)--estimating the population value of $\sigma^2$ from the variance of scores _within_ treatment groups--is exactly the same as for the _t_-test except that we use all three (or more) sets of data instead of only one (as in the preceding examples where we were dealing with a single distribution of differences) or two (as in the _t_-test for impaired or uncorrelated _S_s, as in the calculations given on pp. 23-28 of Chapter 10).

A simple illustration of the analysis of variance as applied to a "randomized blocks" design will be given in Chapter 11.

# Chapter 9
# Procedural Patterns of Instructional Experiments

We have previously differentiated (Chapter 5, page 17 ) between two meanings of the term "design": (1) the broader meaning of the total activity involved in experimental research, from hypothesis formulation through analysis and interpretation of data; and (2) the more specific meaning of experimental structure or structural pattern. The latter is restricted to those formal aspects of an experiment's structure which govern analysis and interpretation, and which therefore affect both the internal validity and the precision of the experiment. Another term often used as a caption in describing experimental arrangements is the term "procedure." This again has both a broader and a narrower meaning. The narrower meaning generally connotes details of procedures (such as the specific instructions given the subject, for example, or the minute-to-minute sequence of events during the course of experimental administration).

The topic of the present chapter concerns procedural aspects of an experiment in a less detailed sense, relating to the major sequence of experimental treatment and measurement. We want to consider here questions that involve the so-called "before-after design" as compared with the so-called "after-only design" (or "post-test only design"). Here we are referring to the sequence of (1) observation or measurement, (2) experimental program treatment, and in some instances (3) a second application of the measuring instrument. To further delimit the present topic, we might term this aspect of experimental design "sequence pattern" to differentiate it from the broader, more inclusive meaning of design, or "procedural pattern"--as distinguished from "structural pattern." These two aspects of experimental design overlap, but a differentiation in meaning appears useful, structural design referring essentially to aspects which stress the sampling and assignment of learners (subjects) to treatments (and the ensuing consequences for the valid and sensitive analysis of whatever measures of instructional effects the experimenter chooses to utilize), whereas "sequence pattern" or "procedural pattern" refers specifically to the sequence of observations or measurements in relation to the timing of the experimental program.

## The "Before/After" logic further examined

As has been emphasized previously, the basic logic of the experimental method determining whether a given treatment had an effect implies a research design which gives measures of the relevant behavior "before" and "after," or "with" and "without," exposure to a program, in such a way as to isolate the source of behavior changes produced. To do this, a variety of experimental design-patterns may be suitable, depending on circumstances. For any design, however, the reader may be reminded that the crucial aspect distinguishing the experimental method from survey or correlational methods is that presentation of the communication (prior to the "after" measurement) is arbitrarily given to or withheld from predesignated samples of people at the choice of the investigator.

Before reviewing this logic and then considering the procedural patterns whereby it may be implemented in an instructional-media experiment, let us digress briefly to consider the requirements for a "before then after" sequence of measurement in the case of experiments in other fields.*

The before measure is not always required in certain types of experiments. Often, of course, it clearly is, with requirements that are closely parallel in the instructional-media experiment and that of the experiment in other fields of inquiry. For example, in the physical-science laboratory, the length of a metal rod has to be obtained both before and after heating it in order to determine the coefficient of linear expansion, and the volume of a gas needs to be determined before and again after subjecting it to change in temperature or pressure in order to obtain measurement in accordance with Charles' Law or Boyle's Law. But in many kinds of experiments, measurement of a before level is not necessarily called for--either because the before level is clearly zero, or because change due to treatments is measured directly rather than being calculated as the difference between before and after status. As an instance of the latter case, in the verification of Hooke's Law in the physics laboratory, deformation or stretch is measured directly starting from zero change (not zero length) as tension on a vertical spring is increased by adding weight. In the former case (zero starting point) the crop yield in agricultural experiments starts in a literal sense "from scratch" (or "from seed").[*]

### "Before" or "no-treatment" measures in instructional experiments

*Assessing the effect of a given treatment*. In an instructional experiment, an initial-level baseline is ordinarily required when we seek to determine the effect of some particular treatment. As noted above, one condition under which a before measure would not be needed lies in instances (e.g., the yield from a farm plot) when the before value is, in some sense at least, clearly zero. This is seldom a reasonable assumption in instructional experiments, though there are exceptions--as in the case of artificial or uncommon learning tasks such as the learning of nonsense syllables or, say, the learning of Sanskrit verbs or Morse code by kindergarten children. The other kind of exception (measurement of change directly) applies even less often, since we are ordinarily not in a position to make direct observation of the learning process itself but only infer learning from measures of "achievement" after an opportunity for learning has been given.

In the general run of experiments on the effects of instructional media on knowledge, skills, attitudes or other curricula objectives, we can by no means assume a zero starting point. Rather, we generally have to assume some undetermined degree of initial skill, knowledge or attitude. Therefore, a measure taken only after an instructional treatment does not indicate the effect of the treatment itself, but merely the after status that can have resulted from some unknown contribution of treatment plus some unknown amount of effect from pre-treatment experience. In other words, when measures are taken after a treatment without an appropriate measurement of "before" status, the effect of treatment and of prior experience are wholly confounded. Thus, in determining the effect of any one treatment--i.e., testing the hypothesis $(X_1) > 0$--we always need a "before" measurement unless we feel warranted in making the assumption that the before level is zero. Though this may seem obvious on reflection, it is nevertheless surprising how often this elementary consideration is overlooked, so that responses after a treatment are tacitly assumed to be "effects" actually produced by the treatment, with no control on whether the same "effects" might have been present *before* the treatment.*

*Comparing the effects of two (or more) instructional treatments.* However, there is another frequent kind of situation in which we do *not* need a before measure--namely the situation in which we are testing a hypothesis concerning which of two treatments produces the *greater* effect. As a matter of fact, there is ordinarily less interest, in educational experimentation, in merely raising the question of whether a particular treatment is more effective than no treatment at all than there is in comparing the effects of two treatments. That is, using the notation introduced in Chapter 7 (p. 7-38), the hypothesis $(X_1) > 0$, is generally of less interest than testing the hypothesis that the effect of $X_1$ is greater than that of $X_2$, or $(X_1) > (X_2)$. Just as we already know that crops will grow somehow, given any reasonable chance, so we know that children will somehow learn something from nearly any form of relevant instruction, thus demonstration of merely this is often fatuous (though this fact does not necessarily seem to be obvious in all cases, as witness the plethora of experiments--cf. Chapter 12--calculated to do no more than simply demonstrate for the $N^{th}$ time that "films can teach" or that "TV can teach").

*After-only comparison of two treatments, $X_1$ and $X_2$.* In a hypothesis-testing experiment comparing the effects of two treatments, $X_1$ and $X_2$ in terms of a dependent variable, $y$, to test a hypothesis about the influence of variable X [without reference to the estimation problem of determining the *size* of $(X_1) - (X_2)$], we need only test whether $y_1 > y_2$, or $y_1 - y_2 > 0$. That is, if $y_0$ is the before measure, we logically do not need to test the second-order difference $(y_1 - y_0) - (y_2 - y_0)$, because in this expression of the logic of the comparison the no-treatment or "before-level" term, $y_0$, drops out. Thus we can compare the *difference* in effect directly, in terms of $y_1 - y_2$, without ever inquiring about the size of $y_1 - y_0$ (the effect of $X_1$) or of $y_2 - y_0$ (the effect of $X_2$) as such. Hence even though a "before" (or "without treatment") level needs to be established (unless it can be assumed to be zero) in order to determine the effect of *any given treatment*, it is *not* needed to test a hypothesis about which of two treatments is the more effective. In other words we may say that in comparing two treatments, assigned at random to two groups of subjects, an *after-only* comparison will suffice to provide a wholly valid experiment.*

*"After-only" terminology in measuring the effect of a single treatment ($X_1$).* It can also be seen that the logic that applies to determining whether one of two treatments has greater effects applies also, in every particular, to the determination of whether a given treatment $X_1$ has *any* effect. The reason for this is that a "no-treatment" control condition, which is needed to assess the question "Is $(X_1) > 0$?" can always be considered in itself a comparison treatment, $X_0$, allotted to a control group. The basic paradigm of the valid "after-only" experimental determination of the effects of a particular treatment is thus to test the hypothesis that the value of the criterion measure, y, for any one treatment $(X_1)$, is greater than that for the no-treatment control $(X_0)$--that is, that $(X_1) > (X_0)$--by comparing the dependent measures $y_1$ and $y_0$ for experimental and control groups that were, respectively, exposed and not exposed to the treatment, and testing the null hypothesis $Y_1 - Y_0 = 0$. Since in this case the no-treatment group serves as a control, or before level, the

$X_1 - X_0$ comparison is sometimes expressed as a comparison of $X - C$, using "C" to stand for "Control." Note however that the "control group" here used is a "no-treatment control" and that this use of the term "control group" is to be carefully distinguished from other forms of control groups required for ruling out other alternative hypotheses. (See also the comment in Chapter 7, p. 34 concerning the special and unfortunate use of the term "control group" to refer to a condition of "conventional instruction" with which an instructional-media program is compared.)

This special sense of valid "after-only" comparison (actually, "after treatment" compared with "after no treatment") has sometimes been confused with the invalid attribution of effects to programs where a necessary no-treatment measure was lacking, or, on the other hand, with the valid only-after-treatment measurement that, as seen above, suffices for comparing the difference in effects of two different treatments ($X_1 - X_2$). The invalidity of the former of these (a single after-only measurement with no comparison condition) contrasts with the potentially complete validity of the comparative "after-only" pattern in which an after measure for $X_1$ is compared either with a no-treatment control condition $X_0$ (to yield a measure of the effect of $X_1$) or with an alternative treatment, $X_2$, to yield a measure of the difference in effects of $X_1$ and $X_2$. It should be emphasized that the treatment-versus-no-treatment comparison is thus merely a special case of the basic paradigm of all first-order or unqualified comparisons in educational experiments: that is, the test of the experimental hypothesis that treatment $X_1$ has a greater effect than any other treatment, $X_2$, includes as a special case the situation where $X_2$ (now, using the shorthand convention, termed "$X_0$") is no treatment, so that instead of testing whether $Y_1 - Y_2 > 0$, we test whether $Y_1 - Y_0 > 0$.

## Relation of "after-only" pattern to "before/after" rationale

It may be helpful to relate the foregoing explicitly to the general rationale of before-after change as previously set forth. Let us recall that, in this rationale, in order to show that a program caused an effect we have to know not only what those who received the program were like after having it, but also what they were like before this experience--or, more precisely, what they would have been like had they not had it. The advantage of the latter formulation is that it implies the necessity of ruling out the possibility of other causal factors entering in to influence those exposed to the program between the time of the "before" and "after" measures if these are obtained at different points in time. This possibility is automatically taken care of in this after-only pattern, where the "before" measure from a no-treatment control group is obtained concurrently with obtaining the "after" measure from the experimental group that was given the treatment.

In this case the status of what those who got the program would have been like without getting it (which is the "before" point that we really want) is obtained without actual before-and-after measurement of the same individuals. The "before" measure, obtained concurrently with the "after" measure for the exposed group, comes from the comparable group of individuals (control group) from whom the program was arbitrarily withheld. "Comparability" (i.e., equivalence except for chance factors--see Chapter 3, p. 30 ff.) is, of course, assured if the determination of which individuals will get the program and which ones will not is determined on a purely random--i.e., chance--basis, so that aside from the one factor of having had the program or not, the only factors that can differentiate the two groups are, in fact, chance factors. Thus, by all odds the simplest, most straightforward, and safest of the major procedural patterns for assessing the effect of a given treatment $X_1$ ,or in hypothesis-testing terminology, for testing the hypothesis $(X_1) > 0$, is *not* before-and-after measurement of the same individuals at different time points, but rather the "after-only" comparison of comparable (randomly assigned) "with" and "without" groups tested at the same time--i.e., after the treatment has been given to the experimental group.

*Possible confusion in the use of term "after-only" when applied to assessing the effects of a single treatment.* Some confusion seems to have been occasioned by what might seem the paradoxical term "after-only" as applied to the measurement of a particular treatment (e.g., a film). The confusion arises simply from the failure of readers to remember that, at least with proper randomization of treatment and no-treatment groups, a before measure (no-treatment measure) *does not have to be obtained from the same individuals* as the after measure, but rather may be obtained from an equivalent or comparable no-treatment "control" group that is tested concurrently with the experimental treatment group. Since the testing for this control group takes place after the parallel experimental group received the treatment, there is the seeming confusion that the "before" (no-treatment) measure is also, temporally, an "after" measure. Looked at the other way around, in this "after-only" design the "after" measure for the no-treatment control group does, as we have seen, provide a "before" measure. It is evidently for this reason that the "after-only" terminology (Hovland, Lumsdaine, and Sheffield, 1949) and "post-test only" designation (Campbell and Stanley, 1962) are sometimes thought to be confusing. If "after-only" is in this sense a misnomer, it appears, however, that we are stuck with the term through wide usage. Clarity may, however, possibly be achieved by avoiding the term "after-only design" which, aside from other previously noted ambiguities, is perhaps too readily confused with the invalid after-only, no-treatment, no-control basis for erroneously trying to assess a program's effects, and by speaking only of after-only comparisons--the term "comparison" hopefully suggesting either the $X_1 - X_2$ or $X_1 - X_o$ comparison, rather than what Campbell and Stanley (1962) call the "one-shot case study" (after measure for only one group with no baseline for comparison).

"Before" or "No-treatment" measures as a special case of control for alternate hypotheses. In terms of the discussion given in Chapter 5, it can be seen that in determining the effect of a given treatment, the before measure can really be considered as a control for the alternate hypothesis that the treatment group had prior knowledge of the subject matter. The fact that in comparing two treatments $X_1$ and $X_2$ this before-level control can be dispensed with (see above) by comparing $y_1 - y_2$ directly--instead of comparing $(y_1 - y_o) - (y_2 - y_o)$-- is parallelled by the fact that similarly, any control for any alternate hypothesis about effects of an extraneous variable, $X_c$, can also be dispensed with in comparing effects of $X_1$, with $X_2$ if we can be assured that the extraneous variable, $X_c$, applies equally to the two treatment conditions $X_1$ and $X_2$. Thus just as we do not need to know the before level of knowledge in comparing $(X_1)$ and $(X_2)$ if only we know (by virtue of random assignment to $X_1$ and $X_2$ that it differs only by chance, so also we do not need to measure the status of the groups with respect to other possibly biasing factors--e.g., learning ability—provided that comparability with respect to such factors is assured by proper randomization.

Procedural basis for the validity of the after-only comparison. The logic of this "after-only" experimental pattern is that of measuring selected items of student behavior (reflecting attitude, skill, knowledge, etc.) with and without the experimental treatment in such a way as to have the "after-non-treatment" measures provide a valid index of what the experimental group would have been like had they not been exposed to the treatment. This, as we have seen, implies some kind of randomization safeguard against selection bias as a way to assure what is commonly called "comparability" or equivalence of groups. That is, we assign subjects in terms of a randomizing procedure to two groups which are solely differentiated, other than in terms of assessable chance factors, by the one particular of whether, prior to measurement, the experimental treatment is given (in the case of the experimental group) or withheld (in the "after-non-treatment" control group). We then determine whether we can reject, at a satisfactory level of confidence, the null hypothesis that the treatment has no effect. Since the probable extent of the differences due to chance can be precisely estimated, a comparison of the obtained difference with those predictable by chance alone can permit a rigorous attribution of observed differences to the effects of the program.

Strictly speaking, this will be the case, however, only if two other conditions are fulfilled. The first is (cf. Chapter 4, p. 16 ) that the random assigment either took place immediately before the administration of the experimental program, or that we can insure, through procedural controls, that conditions subsequent to the random assignment were strictly "comparable," if not identical, for the two groups—with the one exception of the experimental program. The second condition is that either the after measure be taken immediately after the program, or—if this is unfeasible or undesirable (as is often the case, e.g., when the dependent variable of interest is long-term recall or retention), that the conditions subsequent to the program but preceding the after measure were also comparable, or such as not to bias the comparison.

The use of the terms "comparable" or "equivalent" here cannot imply exact identity of conditions other than the experimental variable, for this, as earlier emphasized, is an ideal which in practice is never possible to achieve. The reader may recall it is for this reason that the term "unbiased" was used. What we have to try to insure is that any extraneous influences, other than the experimental variables that operate subsequent to the act of randomization are either so slight as to be of negligible consequence (and we can never be completely sure of this) or, preferably, that all such extraneous influences are also randomized in the sense that two further requirements are met: One, that they act randomly rather than systematically on both of the two measures that will be compared as a test of the program's influence; two, that this random influence is fully taken into account in our estimate of chance error. The implications of these conditions have been noted earlier in our discussions of the null hypothesis and the statistical test by which it may potentially be disproved (Chapters 3, 4, 5, 8 ) and will also enter into our further consideration of validity in Chapter 10 .

The "before-after design"

Although the "after-only" comparison when applied to "after treatment" versus "after no treatment" groups, as just discussed, does actually employ a "before" measure, this before measure is--to reiterate--obtained from a separate control group rather than from the same individuals who were given the experimental treatment. The contrasting terminology used by Hovland, Lumsdaine, and Sheffield, as well as others, for an alternative experimental design is that of the "before-after design." This generally refers to a procedural pattern in which the before measures are obtained from the same individuals who provide the after measures. Though it may seem more logical and reasonable to thus obtain the measures from the same individuals--at least, it may so seem for those who have not grasped the full power of the insurance provided by randomization--this procedural pattern of repeated measures on the same individual in fact introduces complications which render it, for many purposes, much less satisfactory than the after-only pattern.

As we have seen, the advantage of the after-only pattern is that the measure of what the experimental subjects "would have been like" without receiving the experimental treatment can readily be obtained, from a control group, at the same time as the measure of "after" status is obtained from an experimental-treatment group. Thus, extraneous events correlated with passage of time are constant for the two groups. By contrast, in simple "before and after" measurement, with the same individuals before treatment and after receiving an experimental treatment, this can never be the case; there is inevitably a time difference in the two measures, and thus the hazard that events other than the experimental treatment, occurring during this time interval, may affect the after measurement.

Measurement of separate groups at separate time points. Another possible experimental pattern is the obtaining of a before measure from a separate control group at an initial point in time, $t_1$, before giving the treatment to the experimental group, and comparing this with the after measure obtained from the experimental group at the later time point, $t_2$, following the treatment.* However, there is seldom any advantage or justification for this pattern as compared with a straight after-only comparison, since additional "history" and "maturation" variables associated with the difference in the time points then enter and have to be controlled, while potential advantages of before and after paired measures for the same individuals are also lacking. Accordingly, this experimental pattern will not be further considered herein.

Alternate hypotheses for before-to-after changes in a dependent variable. The extrinsic events that intervene between the time points of before and after measurement imply alternate hypotheses about the observed effects that include Campbell and Stanley's categories of "history," "maturation," "testing," "instrumentation," "regression," and "mortality" (see Chapter 6). However classified, they must be ruled out by appropriate control if the attribution of effects is to be unambiguous. It is evident that, at least in the case of any instructional-media treatment that is conducted under naturalistic conditions and that covers an appreciable span of time, this will always require, in principle at least, a control group or comparison group which receives a different treatment (which may be, and often is, "no treatment") during the interval between the before and after measurement.

In particular, the attribution of cause to a single program, $X_1$, on the basis of before-and-after measures for individuals exposed to that treatment will generally require a control group, measured at the same before-and-after time points, to assess the effects of any outside influences, other than the experimental treatment, that may have been present during the interval from before to after. The absolute necessity for such a control group in assessing a single treatment may, in the judgment of the experimenter, vary with circumstances. For example, if the before measurement can realistically be made just before the program, and the purposes of the experiment can be met by giving the after measure just after the program, and if conditions are such--e.g., when the program is the continuous showing of a film--that the subjects are meanwhile insulated from outside sources of influence, the experimenter might decide to dispense with the control group.

In other words, the need for such a control group, in terms of ruling out an alternative hypothesis about the effect of a particular treatment, arises from the fact that other sources of outside influence have occurred. This is particularly likely if the time interval between before and after measures is appreciable and the exposure of the subjects to other influences is not under control during this time. For example, if a before measure for an indoctrination film dealing with a current news topic is given a week before the experimental treatment is administered, it is quite likely that other sources of influence--newspapers, radio, etc., may have caused changes in the meantime. Hence, the difference from before to after cannot be attributed unambiguously to the particular treatment (film) unless there is a corresponding measure of what change *would have occurred from before to after without exposure* to that treatment. We then have a before and an after measure for an experimental group, a before and an after measurement for a non-exposure control group.

In such a case, the measure of effects for the experimental group is the *difference* between the *changes from before to after* in an experimental group minus those occurring in the control or comparison group. Thus, barring the exceptional instance in which the experimenter may decide that other forms of control are adequate, the before-and-after procedure always involves *four* measures instead of the two ($y_1$ and $y_2$) required for the after-only procedural pattern--namely (using the subscripts "B" and "A" to indicate, respectively, the "before" and "after" time points, $y_{1B}$, $y_{2B}$, $y_{1A}$ and $y_{2A}$. Furthermore, since each measure employed is subject to variability or chance errors, the amount of error variance in a comparison involving four measures is greater than one involving only two measures; the disadvantage thus introduced will have to be more than compensated for by other features before the before-after pattern can be preferred to the after-only pattern.

Before discussing these implications further, note as a matter of terminology that the term "before-after" may (like the term "after-only") be a confusing one, because it has two quite distinct meanings. The first of these is the generic meaning of the concept, model or rationale of change from before to after, or status with and without a treatment, which--as we have just seen--can well be implemented by the so-called after-only pattern of experimentation. The second meaning refers to that special from or method of implementing the general before-to-after change rationale in which measurements from the same subjects are used to provide both the before and the after measures for an experimental group. In the present discussion we will use the term "before-after comparison" in reference to this particular procedural pattern--just as we have used the term "after-only comparison" in a special sense to refer to a comparison between two (or more) "after" measures. The abbreviation "AOC" for the "after-only comparison" pattern can thus be contrasted with the "BAC" pattern, where BAC stands for "before-after comparison" with the special implication that before and after measures are obtained for the same individuals before and after $X_1$, and also at the same two time points, for some comparison condition, $X_o$ or $X_2$. As noted above (and as will be elaborated presently), such a "BAC" pattern implies comparison, on the two groups, of the two differences, or changes, from before to after (rather than comparison only of the after measures for two groups). We thus might prefer to think of "BAC" as standing for a "before-to-after-change" pattern.

Paired and unpaired before-after measures. Actually there are two important variants of the procedure of employing before-treatment and after-treatment measures for the same individuals in two groups, $X_1$ and $X_2$. The first of these is where paired measures are used, so that we pair a particular individual's before status with his after status so as to obtain individual change or difference scores for each person, each equalling the difference between after and before status for that specific individual. In general, we shall further limit the abbreviation "BAC" to designate this procedure of obtaining paired before-to-after changes for the same individuals--as distinguished from a procedure in which before and after measures are obtained from the same group of individuals but without identifying or pairing up by individual the particular before and after scores. In the latter case, the differences from before to after are simply differences for an experimental group as a whole rather than showing changes from before to after for specific individuals. Before-After Differences for the same group of persons (but without identification of individual changes) might be called the Before-After Group-Difference pattern, or simply "Before-After Difference" pattern, and could thus be abbreviated as the "BAD" pattern. Such a term might not be wholly inappropriate in its connotation, inasmuch as we shall argue that this is generally a poor form of procedural pattern to adopt in an experiment.

### "B.A.D." patterns with "mortality" losses

When repeated (e.g., before-and-after) measures are made in practical situations it is very often the case that some individuals were not present on one occasion or the other. The available before measures are then for one group of persons and the after measures are for a group that is of somewhat different though overlapping composition. This fact poses a problem in how to analyze the results in such a way as to avoid possible selection artifacts, since we cannot necessarily assume that absences were randomly determined (e.g., absentees may differ in health or, in situations where attendance is not required, in interest). Even were this not the case, moreover, it is hard to deal simply with individual changes unless the same individuals are considered on both occasions, and even if individual changes are not the unit of analysis (as they generally should be to realize the potential advantage of a

"BAC" pattern), the statistical analysis is likely to be messy because of the unequal number of cases ("unequal N's") for the measures to be compared. However, this last fact is less serious, potentially, than the selection bias that may result in consequence of non-random selection factors that affect who was present or was not present at a given time.

Similar problems due to absentee "mortality" in the data may also be encountered in an "AOC" pattern if both more immediate and later measures are made to test "retention" or, more generally, changes in effects of programs with passage of time. When more than two measures are made--as with a before or "pre" measure, an immediate after or "post" measure, and a further "post" measure after an interval--the composition of the group of individuals can be even muddier, containing some individuals who were present only in one of the three occasions, some who were present all three times and some who were present on two of the three occasions--first and second, second and third, or first and third. Of the latter three categories, the last is likely to be the smallest; the other two may be of considerable size, particularly if interval 1--2 and/or interval 2--3 was a long one.

In general, the best course of action in these situations is to restrict the analysis to those individuals who were present on both (or all) occasions.* This would seem, if mortality is appreciable, to have the disadvantage of "sacrificing N" (reducing the number of "cases") to some extent. This disadvantage is generally not serious (unless the mortality rate is so high as to jeopardize the validity of the whole experiment anyhow). Whatever loss in sensitiveness is thus incurred from reduction in "N" is generally more than compensated for by the advantages in a cleaner, less messy analysis as well as by the fact that spurious effects due to differential selection are avoided. Finally, as we have before implied, there is little advantage to a "B.A.D." experiment unless the "BAC" analysis, which requires the individually paired measures that are necessarily limited to those present for both (or all) measures, is employed. This will become apparent through the discussion of the comparative merits of the "BAC" and "AOC" patterns which follows the next section on the logic of before-after comparisons.

Logic of Before-After-Change analysis

Basically, there are only two ways of establishing experimentally that an instructional treatment, $X_1$ --say a film--produced a gain in achievement, or that two treatments, $X_1$ and $X_2$, differ in average effectiveness, for any given population of learners.* One of these ways, utilizing an "after-only comparison" ("AOC"), is straightforward: Assign, randomly, individuals to two groups; give one group $X_1$ and the other group $X_2$; then give each group a test, and compare the scores. (If we are assessing the effects of the single treatment $X_1$, then $X_2$ becomes $X_0$, the no-treatment condition, and thus affords the "before" baseline against which to compare the level achieved by $X_1$.) By comparing the test scores we now determine whether variable X made a difference by asking whether, of two otherwise comparable groups (differing, aside from chance, only in the treatment received), one is reliably ("significantly") higher than the other--that is, more difference than we can reasonably expect as a result of chance. Thus Method One for proving that an experimental variable had an effect is to prove that an experimental and a comparison group (drawn from the same population prior to the experiment, in such a way as to differ only by chance) differed reliably after the application of the experimental variable. Method Two consists in proving, with a before-after comparison ("BAC") that the changes from before to after were reliably greater for $X_1$ than for $X_2$. As Hovland, Lumsdaine, and Sheffield have noted (1949, p.320):

> . . . in an experiment the significance of a before-after change in the experimental group is usually not the important consideration. Rather, the important consideration is the comparison of the changes in the experimental group and the control group. In a before-after study, events other than the experimental variable which intervene between the first and second measurement can produce changes, so that a change in the experimental group may be accompanied by a corresponding change in the control group, indicating that the change in the experimental group was due to factors other than the experimental variable. Thus the test of significance must demonstrate a reliable difference between the changes in the experimental group and in the control group.

Valid and invalid analysis of "BAC" data

It is necessary to stress the requirement of establishing a reliable difference in changes because of the frequent invalid use of other forms of analysis. In discussing these we may use $B_1$ and $B_2$ for the before measures for two treatment groups, and $A_1$ and $A_2$ for the after measures.* Our assertion is that in employing the "BAC" pattern it is necessary for a valid analysis to show that the difference $A_1-B_1$ is reliably greater than the corresponding difference $A_2-B_2$ (i.e., that $A_1-B_1 - [A_2-B_2] > 0$).** Two invalid forms of analysis sometimes employed both suffer from the fallacy that they amount to trying to prove the null hypothesis. The more common of these invalid forms of analysis is to show that $A_1$ and $B_1$

do differ significantly (i.e., reliably), whereas $A_2$ and $B_2$ do not differ reliably--that is, that there was a reliable change for $X_1$ but not for $X_2$ (or $X_o$). However, as noted by Hovland, Lumsdaine, and Sheffield (1949, p. 320):

> . . . it is not sufficient merely to show that a reliable change occurred in [an] experimental group[, $X_1$,] but that a reliable change did not occur in [a comparison] group. The invalidity of such a procedure is illustrated by the absurd case in which a change of 10 per cent in the experimental group is found to be significant but a corresponding change of 9 per cent in the control group did not quite meet the criterion of significance. In such a case it would be absurd to conclude that the 10 per cent change in the experimental group was due to the experimental variable rather than a trend caused by other variables. Even when a zero change or a difference in the opposite direction is found in the control group, it is not sufficient to demonstrate, by the preceding statistical test, that the change in the experimental group was reliable. The reason for this is that a zero difference or a small but unreliable difference in the opposite direction in the control group, while it fails to prove that a change took place in the control, does not prove that no change took place in the control.

Though the invalidity of such a procedure is clear--either in terms of the absurdity that results in extreme cases or in terms of the logical impossibility of proving the null hypothesis--the error is nevertheless a prevalent one. As Campbell and Stanley have pointed out (1962, [illegible].pp. 171-246)

> Use of significance tests presumes but does not prove or supply the comparability of the comparison groups. . . many researchers have computed two ts, one for the pretest-posttest difference in the experimental group, one

for the pretest-posttest gain in the control group. If the former be "statistically significant" and the latter "not," then they have concluded that the X had an effect, without any direct statistical comparison of the experimental and control groups. Often the conditions have been such that had such tests been made, the difference would not have been significant (as in the case where the significance values are borderline, with the control group showing a gain almost reaching significance). Windle (1954) and Cantor (1956) have shown how frequent this error is.*

The other invalid form of analysis is to show that whereas the "after" difference $a_1 - a_2$ is reliable, the "before" difference $b_1 - b_2$ is not reliable. If the experiment is otherwise valid in terms of properly randomized assignment, just the first of these two demonstrations (i.e., that $a_1 - a_2 > 0$) is sufficient, of course, as an "AOC" comparison; in this case, showing that $b_1 - b_2$ is not a reliable difference should be redundant. On the other hand, if the randomization conditions for a valid experiment were not met, showing that $b_1 - b_2$ is not a reliable difference does not suffice to establish comparability, for two reasons: (1) the null hypothesis cannot be proved, and (2) even if $b_1$ could be shown to be exactly equivalent to $b_2$, the establishment of initial equality in terms of the prior knowledge measured by the before test would not show that the groups were comparable in other respects. Once again, we return to the basic principle that "comparability" must be established, for a true experiment, by the procedure of randomization employed, and not by post-hoc statistical analysis of indices of "comparability."

## Problems and Relative Advantages of "BAC" and "AOC" Patterns

A useful discussion and analysis of differences and comparative advantages of the so-called "before-after design" as contrasted with the so-called "after-only design" has been given in Appendix "C" of the Hovland, Lumsdaine, and Sheffield volume (1949). Although this remains an excellent treatment, it has some disadvantages for the present purpose, in that the discussion of the treatment is--though in the main more generally applicable to a wider class of situations--presented in the specific context of studies of the effects of military indoctrination films where effects generally were measured in terms of the percentages subscribing to attitude or opinion statements. Thus the relevance of the main argument to other forms of measures and other kinds of instructional instruments has to be extrapolated and disentangled from the more particularistic aspect of this presentation.

Hovland, Lumsdaine, and Sheffield distinguished three major kinds of considerations that pertain to the relative merits of "before-after" and "after-only" comparisons--namely: (1) statistical and sampling factors (or we could say questions of internal validity and precision); (2) types of analysis that can be made with the data (or forms of hypotheses which can be tested); and (3) effects of the measurement process on the measurements obtained (which affect the validity of the dependent-variable measures and hence the validity or generalizability of experimental findings). We shall consider each of these classes of considerations, drawing on the Hovland, Lumsdaine, and Sheffield discussion wherever appropriate.

The first class of factors--statistical and sampling factors--can be considered either in relation to the validity of an experiment or in relation to its precision. As we have seen, however, the two patterns, "BAC" and "AOC," do not differ with respect to validity in the case of a true experiment--either pattern can be wholly valid with appropriate analysis. The advantage of the "BAC" form in terms of validity thus applies only insofar as this form can provide partial insurance against bias in the case of "quasi-experiments"--see Chapter 10 (or 15)-where comparability between control and experimental groups and, correspondingly, a rigorous basis for estimating experimental error has not been properly assured by randomization procedures. Our present discussion of statistical and sampling factors will thus relate solely to the comparative advantages of the two patterns in terms of their precision. The main difference arises from the fact that the before-after comparison permits taking the correlation between the paired before and after responces of individuals into account in estimating experimental error. Differences between the two patterns with respect to types of analyses that can be performed also arise from the fact that the before-after procedure permits analysis of individual changes, and thus affords a basis for comparison of effects among subgroups with differing initial characteristics. The third set of considerations--effects of the measuring process on the measurements obtained--involves the possibility that the effects on the learner of making the before measurements (in the "BAC" pattern) may in themselves bias the results obtained in the after measurement. This can be considered as a potentially important aspect of a wider set of considerations relating to the practical problems of valid and efficient measurement encountered in the use of the two procedural patterns.

Relative precision of "BAC" and "AOC" patterns

One advantage of obtaining a before measure is that, particularly if obtained prior to the assignment of subjects to experimental groups, the before measure is one of the indices that can be used in stratifying or pairing the $X_1$ and $X_2$ subjects. This kind of procedure, in which Ss are then randomly assigned within strata or pairs, is one of the ways in which we can achieve a more efficient structural design, with consequent reduction in experimental error. Use of pre-experimental stratifying or pairing techniques is more fully discussed in Chapter 11; however, it may be remarked here that the use of a "before" score as a basis for pairing or stratification is only one of the possibly useful bases that may be employed. For the present, we shall not consider the use of the before scores for this purpose, and shall assume that the test scores of Ss in the $X_1$ and $X_2$ groups are not paired up across groups in either the AOC or BAC patterns. The pairing of scores that we shall deal with here is that of the corresponding before and after scores of the same individual in the BAC pattern.

As stated above, this aspect of the before-after comparison permits the use of correlation of before and after responses in determining the statistical reliability of the differences by which the effects of the experimental variable are to be assessed.* The superiority in precision of the BAC over the AOC pattern that may sometimes be realized (but which is not immutable by any means) arises solely from this correlation. To see the conditions under which this advantage may materialize, we have to consider the basis on which the sampling error for quantative scores can be properly estimated.

In the following discussion the symbol $\sigma^2$, without subscript, will be used to refer to the sampling variance of the group measures used in the comparison of experimental groups. We shall discuss these measures as being experimental group means, as would most often be the case; however, they would also be percentages or other measures. The present notation should be differentiated from other usages in which $\sigma$ represents the standard deviation (S.D.) of a population or of a sample. Thus, for the present purpose $\sigma$ (corresponding in the usual notation to $\sigma_M$) is the so-called "standard error" (S.E.) of a sample mean (which is equal to $\frac{S.D.}{\sqrt{N}}$, where S.D. is the population standard deviation, or an unbiased estimate of it), and N is the number of cases in the sample. In the present discussion the reader may also assume that the Ns for the $X_1$ and $X_2$ samples are equal, though this is in general not a necessary assumption. The comparison in precision between AOC and BAC patterns may now be stated in terms of a comparison of the standard errors of the relevant differences that represent the effects of an experimental variable--namely, of the first-order difference $A_1 - A_2$ for the AOC analysis and of the second-order difference $A_1-B_1 - (A_2-B_2)$ for the BAC analysis. For simplicity of notation we may also use the single letter D to represent before-after differences (thus $D_1 = A_1 - B_1$ and $D_2 = A_2 - B_2$), both in referring to mean differences. The term $\sigma_D$ or $\sigma_{1-2}$ will refer to the standard errors, or square roots of the sampling variances of these mean differences.

In an after-only comparison (or a comparison using only the after measurements in a before-after-pattern experiment) there are thus only two measures, $A_1$ and $A_2$ to be considered, and we need the estimate of the variance of the difference between means of independent samples, namely,

$$\sigma^2_{A_1-A_2} = \sigma^2_{A_1} + \sigma^2_{A_2},$$

where $\sigma_{A_1}$ and $\sigma_{A_2}$ are the standard errors of the after means for Experimental Group 1 and Experimental Group 2 (it being understood that Group 2 may be a no-treatment control group). A statistical test employing "Student's" $t$ can then be used to determine whether the two groups, after the application of the experimental treatments, differ reliably in their responses. Note, however, that on the null hypothesis the distributions of scores in the two experimental groups are samples from the same population, and the best estimate of the S.E. of each mean should be based on one estimate of population variance, derived from the variability shown in both of the samples, as is accomplished in the computation of $\sigma^2_{1-2}$ for $t$.[†]

Thus, the values of $\sigma_1$ and $\sigma_2$ will differ only if the sample sizes, $N_1$ and $N_2$, for Group 1 and Group 2, differ; if $N_1 = N_2$,

$$\sigma^2_{1-2} = 2\,\sigma^2, \text{ or } \sigma_{1-2} = \sigma\sqrt{2},$$

where the estimate of $\sigma$ is based on both samples.

The statistical test needed for the "BAC" analysis, requiring us to determine the reliability of the difference between the experimental and the control changes from before to after, is somewhat more complicated. Since we have assumed random sampling, the measures for two samples for $X_1$ and $X_2$ are uncorrelated. The sampling variance of the difference between differences should thus be estimated as the difference between the two before-after differences, $D_1$ and $D_2$, and can therefore be written:

$$\sigma^2_{D_1-D_2} = \sigma^2_{D_1} + \sigma^2_{D_2}.$$

On the null hypothesis the estimate of $\sigma^2_D$ should in each case be based on the variance of the paired before-after difference scores for both of the two experimental groups following the same logic as above.*

In using a BAC pattern, a $\sigma_D$ would normally be computed from the actual distribution of paired difference before-after scores for each individual, thus taking any correlation of before and after scores into account automatically, without actually computing the correlation coefficient, $r_{BA}$. However, in comparing the precision of the BAC and AOC analysis it is convenient to consider the variance of this before-after difference in terms of the expression

$$\sigma^2_D = \sigma^2_B + \sigma^2_A - 2r\,\sigma_A\,\sigma_B,$$

where $r$ is the correlation between before and after scores. This expression is of course similar for both $\sigma^2_{D_1}$ and $\sigma^2_{D_2}$, and the $\sigma^2_D$ is exactly the same as would be obtained from the distribution of before-after difference scores. In these terms, the sampling variance of the difference between differences is thus

$$\begin{aligned}\sigma^2_{D_1-D_2} &= \sigma^2_{D_1} + \sigma^2_{D_2}\\ &= \sigma^2_{B_1} + \sigma^2_{A_1} - 2r\,\sigma_{A_1}\,\sigma_{B_1} + \sigma^2_{B_2} + \sigma^2_{A_2} - 2r\,\sigma_{A_2}\,\sigma_{B_2}\\ &= \sigma^2_{A_1} + \sigma^2_{A_2} + \sigma^2_{B_1} + \sigma^2_{B_2} - 2r(\sigma_{A_1}\,\sigma_{B_1} + \sigma_{A_2}\,\sigma_{B_2}).\end{aligned}$$

Now if $r = 0$, as would of course be the case with *unpaired* before and after scores in what we called the "B.A.D." pattern, it is evident that

$\sigma^2_{D_1-D_2}$ will always be greater than $\sigma_{A_1-A_2}$, the S.E. for the AOC test, since the latter is, exactly, the first two terms of the last of the above expressions. That is, the sampling variance for the B.A.D. pattern will always be greater than for the AOC pattern by the amount of variance $(\sigma^2_{B_1} + \sigma^2_{B_2})$ contributed by the before scores. Thus the use of the before scores will only increase error, or decrease precision, unless there is a correlation between before and after scores that is taken into account. It is for this reason that the "B.A.D." design is, in fact, bad, from the standpoint of precision. However, with paired before-after scores, we evidently will still not surpass the sensitivity of the AOC comparison unless the correlation between before and after scores is sufficient so that the variance contributed by the second two terms of the expression is more than offset by the deduction afforded by the correlational term.

To see what degree of correlation is required to obtain equal precision for the AOC and BAC patterns, we may set the difference between $\sigma^2_{D_1-D_2}$ and $\sigma^2_{A_1-A_2}$ equal to zero and solve for r. This gives

$$r = \frac{\sigma^2_{B_1} + \sigma^2_{B_2}}{2(\sigma_{A_1}\sigma_{B_1} + \sigma_{A_2}\sigma_{B_2})}$$

In not infrequent cases where the $\sigma_A$ and $\sigma_B$ for each group do not differ greatly from each other, and where there are equal Ns in the two groups, all the $\sigma$'s in the above can all be replaced by the same value, $\sigma$, giving: $r = \frac{2\sigma^2}{2(2\sigma^2)} = \frac{1}{2}$. Thus the correlation between before and after scores must, in this case, exceed .50 before the Before/After comparison will surpass the After-Only comparison in precision.*

Often--especially if the before scores are at a low level and the gains from the experimental program are sizeable--r will not be this high. In such cases the after-only form of comparison is not only quite as valid as the before-after form of comparison but is more precise. Finally, even where the B/A correlation is high enough to afford a real increase in precision by adopting a BAC in preference to an ACC pattern, the experimenter should also, before adopting it, consider alternative ways in which greater precision might also be obtained--such as increasing the size of the experiment somewhat, or improving the sensitivity of the test instrument used to measure the criteri[illegible]ependent var

### Special problems arising from the use of the before measure ("BAC" patte

The "BAC" pattern implies repeated measurement of the same individuals--at least twice, once before and once after the experimental programs. If, in addition, measures of delayed as well as immediate retention after the program are desired to test implications of the experimental hypothesis, then there may be occasion for still a third measure (unless separate sub-groups are used for the short term and long term "after" measures). In addition to the "mortality" difficulties noted earlier, several problems arise as a result of this need for repeat testing of individuals. Some of these problems are particularly bothersome in the case of studies where opinion or attitude changes result instruction are the focus of interest; others are of concern in any instructional experiment.

In general, these problems represent disadvantages of the "BAC" pattern's requirement for the additional act of measurement necessary to secure the before measure. It may or may not be easy in a given experiment to overcome these disadvantages, or, even if so, to do so without sacrificing or compromising other desirable aspects of the total experimental design. We want here to identify some of the main problems and also to suggest some ways in which they can sometimes be resolved. However, their existence further suggests that the proposal to use the "BAC" pattern should be sufficiently justified--either on grounds of a needed increase in precision not more easily obtainable otherwise, or because the analysis needed for testing an experimental hypothesis requires paired before/after measures for individual subjects (see pp. 43 ff).

The special problems to be considered include effects of taking a first (before) test on (1) reaction of subjects to the second (after) test, (2) on their reaction to the experimental variable, as such, and (3) on their perception of the total experimental situation--and thus, indirectly, also on one or both of the first two elements. [These three areas of problems correspond roughly to the third, ninth, and eleventh of the factors listed by Campbell and Stanley (1962), which were discussed in Chapter 6 in the context of alternate hypotheses concerning the outcomes observed in an experiment.] In addition, we should consider certain practical matters concerning the expenditure of research resources in relation to the requirements for giving a test before as well as after the experimental program. These latter considerations, which are sometimes of considerable consequence to the economy or even the feasibility of an experimental arrangement, primarily affect the relations of the experimenter to his budget and to the teachers and administrative personnel with whom the practical arrangements for an experiment must be worked out, whereas the first three problems may affect the validity or sensitiveness of the experiment in terms of their direct influence on the behavior of the subjects.

In considering such factors--those that may directly influence the subjects' behavior--we are concerned with the effect of the measurement process on the phenomena under investigation (cf. Chapter 8, pp. ). Here it is desirable, in each case, to consider not only whether a factor is likely to have some appreciable influence on the *S*s' behavior, but also whether--*if* it does--it will have a serious adverse effect on the sensitiveness or validity of the experiment. In some cases, using the before test could have a quite sizeable effect (for example, on the level of the after scores, as compared to what they would be had no before test been given) but without this affecting *differentially* the subjects in the different experimental groups--that is, without interacting with the experimental variable. In such a case the validity and perhaps even the sensitiveness of the experiment as a test of the experimental hypothesis might not be in any way impaired, even though giving the before test did have a demonstrable influence, one way or another, on the average level of the after scores.

*Direct effect of a first test on responses to a second test.* It is known that performance in a second test may be affected by having taken a previous test. This is, of course, the main reason why educational and psychological test construction agencies construct "parallel forms" of tests that are to be widely used for assessing individual achievement or ability.

A subject may, first of all, learn something about the subject matter just from taking a factual test. A similar gain on retesting may occur on other kinds of tests; for example, a test of a psychomotor skill may produce a "practice effect" that represents learning which took place during the test performance itself. Further, particularly if a subject expects to be retested, he may look up some of the answers to questions, on a verbal test, which he did not know or was not sure of. Finally, changes in performance might result, particularly over a short span of time, simply from a tendency to avoid answers given previously.* On the other hand a similar possibility, diametrically opposed to these tendencies to change, has been characterized by Hovland, Lumsdaine, and Sheffield (1949, p. 310) as a "consistency" reaction. This is the possibility that, for certain kinds of subject matter at least, the before-after procedure will stimulate the subjects to give the same response the second time, so that changes indicating the effects of variables are minimized.

The extent of such effects of prior testing might depend somewhat on the rationale given for a second test administration, and on the time intervening between the two, as suggested by Hovland, Lumsdaine, & Sheffield (1949, pp. 310-311).*$

What effect would these possible reactions have on the validity and sensitiveness of an experiment in comparing two treatments $X_1$ and $X_2$? The effects of learning from the test or of looking up answers would of course tend to produce spurious increases in the after measures $A_1$ and $A_2$, thereby perhaps decreasing the experiment's sensitiveness through the introduction of ceiling effects or reducing the amount that $X_1$ or $X_2$ could teach the subjects, thereby tending to reduce the size of the difference $A_1 - A_2$. Either of these effects could of course be reduced or eliminated by using two parallel test forms, one for before and one for after measures (see further discussion below).

The consistency reaction or abundance of repetition might perhaps do likewise and in addition would probably increase the amount of chance error. On the other hand, the *validity* of the experiment would presumably be affected only if these tendencies interacted differentially with the experimental variable, tending to augment the effect of the more effective treatment so that it produced effects that would not occur in a normal application situation where the before test was absent.

*Effects of the before test on student's reaction to the experimental variable.* There is ample evidence that a before test can increase or change qualitatively what is learned from an instructional program. This effect may be a general "sensitization" phenomenon, or may involve a selective "attention-focusing" mechanism. Examples of the latter are found in studies reported by May and Lumsdaine (1958) and Lumsdaine, Sulzer, and Kopstein (1951, 1961) summarized in Chapter 15 (p.        ). It should be noted that the enhancing effect of a pre-instruction test is not necessarily to be viewed as a methodological artifact, but rather may be an instructional (and hence experimental) variable in its own right, as where the pre-instruction test is used as an instructional procedure for the express purpose of increasing the amount or kind of material learned from a subsequent program. Such effects of the before test would be undesirable in an experiment only if their interaction with another experimental variable, which was the focus of inquiry in the study, were to distort (enhance or differentially dampen) the effect of that variable.

*"Sensitization" artifacts.* A special aspect of such a spurious "sensitization" effect has been a particular source of concern in studies in which the effect of an experimental treatment, $T_1$, such as an indoctrination film, was assessed in comparison with a no-treatment condition, $T_o$,

in terms of the effect of the film (or other program) on attitudes or opinions. In such cases there is a realistic source of worry about the validity of the results which may stem from the possibility that a before-test questionnaire may "sensitize" the subjects to the topics to be covered in a program so that their immediate reaction to the program may be spuriously altered. Thus, the before test may arouse an atypical degree of interest in the topic as well as focusing attention on certain aspects of it.[4] A related possibility is that the combination of the before questionnaire and the program will suggest to the subjects, at the time that they are being given the program, that they are being "guinea-pigged," so that the material will be reacted to atypically at that time as well as later when an after test or questionnaire is being answered. This is a prime example of Campbell and Stanley's factor of "interaction of testing with the experimental variable," as mentioned in Chapter 6 (p. 6-32 ).

In experiments on attitude or opinion change or in other cases where it would seem that a "guinea pig" reaction (cf. Chapter 8, p. ) would seriously invalidate the measures obtained in the experiment, it should be noted that a before-after comparison procedure would generally be more likely to introduce this distortion of making the subjects suspicous that an experiment is in progress than would an after-only comparison. Thus, the administration of two successive tests or questionnaires on much the same content, with a relevant program given to them during the interval, would be expected to arouse more suspicion than a single test or questionnaire that came after the program.

It is evident that such suspicion on the part of subjects could, in an experiment on opinion or attitude change on a "sensitive" topic, lead to serious falsification which could wholly invalidate the meaning of the subjects' responses. In such experiments, measurement without awareness on the part of the subjects that their reactions to a program are being assessed would appear essential to valid results.[**] Precautions of the sort taken by Hovland, Lumsdaine, and Sheffield[*] (1949 ) to camouflage and rationalize the purpose of the experiment would thus seem of particular importance in this kind of experiment, but may to a lesser degree also be of concern even where the more ordinary sorts of instructional effects on knowledge or skill are being measured.

"Pre-test frustration." A different kind of undesirable effect may cause difficulty if a long, comprehensive test (such as may be required as a suitable after test) is first given also as a before test on a topic with which Se are initially unfamiliar. The evident stupidity of being required to answer a long list of questions with little basis other than blind guessing for doing so can be so frustrating that it may lead to antipathy or disinterest when the subsequent after-test is given or may in other ways seriously jeopardize subsequent cooperation of students and their teachers in an experiment. This can, of course, also be a problem in an AOC design where a no-treatment control measure is desired. Perhaps the best solution, where a before measure is really required, lies in the use of partial sub-forms of the criterion test. Use of such forms (see below) can also help to reduce time requirements in an experiment and to minimize some of the possible spurious effects of sensitization and other factors discussed above.

Time requirements. Even where the nature of a before-test is such, in relation to students' initial level of knowledge, that it could be readily tolerated and appear reasonable to them, the expenditure of time and arrangements necessary for giving it have to be weighed as a further disadvantage in deciding to use the before-test procedure.

Many experiments on instructional programs must be fitted into a school schedule. In an experiment using a very short program it might be feasible to give before test, program and after test all within a single class session. Usually, however, this would not be the case. With a somewhat longer program it may be feasible to give an experimental program and after test (without a before test) in one class session, or the program in one session and the after test in the next, whereas the giving of a before test would require arranging for the use of an additional (previous) class session.

Any difficulty of getting cooperation from a school may increase rapdily with the number of visits and disruptions of normal schedule that must be made. Also, receptivity to interesting new instructional materials (films, programed books, etc.) may well be diminished when they include the introductio of "one more test." The physical and administrative problems in giving a before test may be further increased if it precedes the experimental program by a considerable period of time, as might be considered necessary to minimize "sensitization" or "guinea pig" reactions (see above).

Resistance to the arrangements for an experiment may also be increased, in many school systems, in an $X_1$ versus $X_0$ comparison where no program is offered to the control classes. In an after-only pattern, this can sometimes be overcome by giving the program ($X_1$) to the control classes after obtaining the $A_0$ measure that follows the giving of $X_1$ to the experimental classes; giving $X_1$ as a "goody" to the control classes does not enter into the experimental test as such but may serve to make the testing procedure more acceptable to a school than if a program is "unfairly" withheld from the control classes. But in any case the giving of *two* tests (a "before" and an "after") with no intervening program will generally present greater difficulties in terms of being made to appear a reasonable and acceptable procedure. These considerations once more suggest that the decision to use a before-after pattern should be based on real advantages that clearly outweigh the greater difficulties it introduces.

*Short-form before tests.* One way in which several of the above-noted difficulties can sometimes be minimized when it is decided that a before test must be used is to use a short form of the criterion instrument as the before test.* Ways of doing this best will depend on the specific purposes and needs of the experiment. In any case using a quite short before test may facilitate administrative problems by taking less time, may reduce or eliminate student boredom or frustration at being tested on an unfamiliar subject matter, and can minimise other undesired side effects such as effects noted above, which can result from testing.*

If the purpose of the before test is merely to stratify subjects to increase precision by segregating subgroups, a very short test, and one not necessarily "parallel" to the after criterion test, may suffice. On the other hand, if gain scores or individual change scores are to be used, it would generally be best to insure that the items in the before test are a truly representative sample of those in the after test. This may be assured by splitting either the entire after test, or each of several sub-test segments of it (each covering different aspects of the subject matter, or different levels of difficulty), into random halves, thirds, or smaller sub-fractions. For example, we might divide the 12 test items, written to cover each of three sub-topics to be presented by a program, into four random sub-tests of 3 items each, making four ¼-length short forms of only nine items apiece.

Any one of these short forms could then be used by all subjects for the before test. On the other hand, we could also assign the four short forms randomly to *Ss* within each experimental group. This would tend to balance out any inequalities in a particular random selection of items and would also afford a before measure, if desired (for purposes of diagnostic program evaluation, for example) on all of the 36 items in the entire test. Though the before level for each item would be based on only ¼ the number of subjects available for the corresponding after level on each item, the degree of accuracy obtained in a fairly sizable experiment might be quite adequate for the use to which the before measure would be put, even though the more precise measure based on the larger number of students might be needed for the after test data.*

However, if other considerations (e.g., pre-program testing frustrations where the *Ss* are ignorant of the topic) did not dictate the use of a shorter before than after test, it might be of advantage—e.g., from the standpoint of maximum reliability of individual change scores—to divide the testing time evenly between before and after tests, using, say, a half-length test both for the before and for the after test for each individual. In this case, the experimenter might or might not wish to introduce the additional complication of seeing to it that each subject got different forms—or the same form—on the two testing occasions; often, it would appear quite satisfactory to let this be randomly determined, in which case, by chance, about half the *Ss* would be receiving the same one (if two alternate test forms, before and after, are used), and half would receive a different form the second time.

## Types of Analysis Possible Only with Paired Before-After Measures

Aside from the possible (though, as we have seen, far from certain) advantage in terms of precision, the main advantage of the before-after comparison lies in its permitting certain kinds of analysis not possible with an after-only comparison. These advantages have two main aspects: (1) the types of measure of effect that can be employed and (2) the kind of hypothesis that can be tested.

The before-after procedure has the advantage that it permits analysis of individual changes. This allows the use of measures which are based on such changes and also allows us to determine effects for various subgroups that differ in initial level of knowledge, skill, attitude, etc. By contrast, a major drawback of the after-only procedure is the limitation it sets on the detail with which changes in response can be described. The measures that can be obtained with the after-only procedure are limited almost to the description of marginal frequency in changes--that is, net shifts in the distribution of responses. For example, consider the following distributions of scores for an after-only comparison of an experimental treatment $X_1$ and a comparison, no-treatment condition, $X_o$:

| Scores: | | 0 | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|---|---|
| number of individuals getting each score | $X_1$ | 0 | 5 | 25 | 40 | 25 | 5 | 100 |
| | $X_o$ | 15 | 20 | 30 | 20 | 15 | 0 | 100 |
| Difference ($X_1 - X_o$) | | -15 | -15 | -5 | 20 | 10 | 5 | |

Taking $X_o$ as an estimate of what the $X_1$ *S*s would have been like without getting $X_1$ permits only an incomplete description of the changes that took place. We know the net change in each score-category, but we do not know what scores the individuals in, say, the "3" category would have received without the treatment.[*] Merely from the marginal-frequency distributions, we would not know this even if the second line of scores were the before scores of the same individuals from whence the $X_1$ scores were later obtained. For example, of the 40 individuals getting a score of "3" after receiving $X_1$, some probably also got (or would have got) a score of 3 on a before test, while others got lower scores; but we don't know what particular scores--some might have moved up from zero to 3, others from 1 to 3, etc. Also, some of these final "3's" might initially have got higher scores, and have gone down (on the second test). The differences in marginal frequency obtained do set some limits on the possible shifts that could have occurred, and comparison of the two frequency distributions does tell us more than merely a statement of the two means (the mean shift in the above example being obviously a change of 1.0, from 2.0 for $X_o$ to 3.0 for $X_1$), or even than a statement of the means and variabilities. But any given pattern of before and after (or $X_o$ and $X_1$) distributions--and likewise any given set of means and standard deviations--is compatible with a variety of possible patterns of underlying shifts. Thus a given size moderate mean shift may result from a large number of small individual gains, or from a smaller number of larger gains.

For some purposes it may be more important to know how many persons improved, or how many gained more than some specified amount, than merely to know the mean shift. A mean shift, or shift in marginal frequencies, may "hide" a variety of different patterns of internal shifts--for example, a cancellation of some increases and some decreases in scores. In addition, the scores may be of such a kind that a few large shifts are not indicative of improvement in the same sense that a larger number of small shifts would be. For this reason, in some experiments [see for example the study of "one-sided" versus "two-sided" programs reported in Chapter 8 of Hovland, Lumsdaine, and Sheffield (1949 ) where the "scores" were subjects' estimates of the duration of the war] the most defensible measure could be simply the net percentage increasing their scores, regardless of amount. Such a measure can be obtained only where individuals' paired before and after responses can be tabulated; it cannot be obtained from an after-only design.

Initial responses as a "sorting" variable

A related circumstance that requires a comparison of paired before-after responses is the desire to test a hypothesis about the relation of change to initial response level. For example, in the Hovland, Lumsdaine, and Sheffield analysis of the effects of a wartime indoctrination film favorable to the British, the question arose as to who was more affected by the film, those already pro-British or those initially anti-British. With a before-after procedure, this question was readily answered by sorting the men into subgroups according to initial response and determining the proportion of change in opinion in these subgroups among film subjects and among control subjects. The difference between film and control changes in these subgroups then gave the effect of the film on each subgroup. With an after-only procedure, however, there would have been no valid way of analyzing effects differentially for those with differing initial levels of opinion. The reason is that the subjects who answered positively after the film would be a mixture of those who initially would have given the positive response and those who would have been negative but were positively affected by the film.*

In general, the test of any contingent hypothesis that predicts different effects of a program as a consequence of differences in initial response can be made only with before-after comparison that permits the pairing of each individuals' "before" response with his "after" response, at least if the initial response is to the same test or test-question as the after response. The only kind of differential analysis in terms of initial status that is possible with an after-only comparison is one in terms of a response which is not changed by the program--for example, I.Q., age or some other stable characteristic. In this case, the analysis can be made by comparing the corresponding subgroups of the $X_1$ and $X_o$ (or $X_2$) groups on the basis of such a differentiating characteristic--e.g., comparing the low I.Q. Ss in the $X_1$ group with those similarly identified in the $X_o$ group, and likewise for the high-I.Q. Ss.[*] But one is not able, with an "AOC" experimental pattern, to make a seperate analysis for those in the $X_1$ group who would have had initially low scores, because there is no way, without before measurement for these same individuals, to identify after the fact what a given individual's initial position would have been.[*]

However, it should be noted also that the "BAC" analysis to test hypotheses that predict greater effects for initial low scorers than for initial high-scorers always also requires a parallel analysis of change for a comparison group as a control on regression artifacts, even though the experimental situation is such that the experimenter might otherwise feel justified in dispensing with a control group to assess the effects of external ("history") or other concomitant influences. The reason for this lies in the inevitability of regression shifts in any repeated measures; see further discussion in Chapter .

## Comparative Advantages of Before-After and After-Only Comparison Patterns

From the various considerations that have been discussed, it is apparent that neither the before-after nor the after-only procedure has all the advantages or disadvantages. In a true experiment there is generally little point in unpaired before and after measures (although they may have some function in a quasi-experiment in which the experimental subjects cannot be assigned at random; in such a case the risk of spurious "effects" which may merely reflect pre-treatment differences is somewhat less with the before-after procedure because it at least takes into account differences in initial level on the criterion variable).

The main advantage of the paired before-after score comparison is the more complete description of effects that it provides, and the fact that it permits testing contingent hypotheses about the relation of program effects to initial status. Often its main disadvantage is the greater opportunity for the measuring process to bias the results obtained: however, this disadvantage can often be minimized with the proper methodology if the experimenter is sensitive to the potential biases and uses ingenuity and care in trying to avoid them. Contrary to prevalent belief, the use of paired before-after scores to assess overall effects in an experiment is not more valid and, further, by no means guarantees greater sensitivity than could be obtained with an after-only pattern. Since the BAC pattern requires comparison of second-order differences between experimental groups, involving four measures (as against only two for the AOC pattern), the relative sensitivity depends on the extent to which the increased variability of the error introduced by the variability of the before scores is offset by reduction in the error term resulting from the correlation between before and after scores. Only if this correlation is high enough (0.5 or better in the simplest case) will the BAC procedure have more precision. Even then it may not be more sensitive if, for example, the size of effects is reduced through the giving of the before test.

Individuals' "before" scores are never absolutely needed unless specified as a qualifying variable by hypothesis: Even though we originally characterized experiments in terms of a "before-after" rationale, we never really need before-treatment scores to establish first-order effects.

The seeming paradox is only an apparent one, as can be seen by considering the two main questions we may wish to answer: (1) Did $X_1$ have an effect? and (2) Was $X_1 > X_2$? In the first instance, given randomization, we need a "before" or "without" level, but can get it from an $X_o$ group. In the second case, we do not even need a before level to test $X_1 - X_2 > 0$. The before level in the latter case would be of service (aside from its possible utility in improving precision) only if we also wanted to estimate the size of the gain for $X_1$ and $X_2$ separately; even here we could (and often best can) get this from a third, no-treatment group, $X_o$.

The before score for identified individuals whose after scores we are comparing is really needed only for testing hypotheses which specify gain as such. This can be either (1) because the hypothesis specifies a measure of effect, such as percent increasing their scores, that is based on individual changes, or (2) because the hypothesis specifies effects differentially as a function of initial level or of some related variable that can be affected by the experimental treatment.

# Chapter 10

# Experimental Structure and Sampling

In Chapter 9 we have dealt with procedural patterns (sometimes called "designs") that involve sequences of treatment and measurement—including the "after-only" and "before-after" sequences. These aspects of procedural sequence in treatment and measurement are not only to be differentiated from the wider, more inclusive meaning of "design" as the totality of experimental method, but also need to be contrasted with the commonest meaning of "design" in the more limited sense. This meaning, as it figures in textbooks on "the design of experiments" and in the sections of research reports captioned "Design," deals with the formal structure of an experiment. Following Tukey (1961), we have chosen to demarcate this as the field of *experimental structure* or experimental "pattern" (cf. p. 5-7 ).

In the present chapter and the next one we return to this aspect of design—that is, the *formal structure of experimental arrangements*, with particular reference to its bearing on the validity and precision of experiments. This question intimately involves the *sampling or assignment* of subjects and/or task-conditions in terms of groups or subgroups, through the use of randomizing procedures—and thereby governs, in large measure, both the validity and the precision of any instructional-media experiment. We shall need to deal here with several basic features of structure, with the experimental operations these imply, and with the influence of these operations on the evaluative qualities that they affect. However, we shall not deal as such with all or even most of the specific structural patterns which are found in the textbooks on "experimental design;" instead we shall limit our attention to what seem to be important features that enter recurrently into a few of the more basic designs which seem primarily useful in experiments on instructional-media. *

Among the principal features or operations and associated underlying concepts to be considered are those of sampling; stratification and replication; matching, pairing and "equating"; and related aspects of procedural and statistical control; factorial structure; and further aspects of randomization as it relates to these. The influence of these operational features on resulting qualities of validity and precision will be noted in this chapter, after which (in Chapter 11) we shall review other factors bearing on these qualities as well as on sensitiveness, efficiency, and generalizability or representativeness.

What has been termed the "internal validity" of an experiment depends primarily on control of bias, and on the ability to make a probability statement that derives directly from one or more aspects of sampling in the assignment of subjects, treatments and experimental conditions. External validity and generalizability, aside from the crucial respects in which these are determined by the definition and implementation of the independent and dependent variables, also derive from considerations of sampling. All aspects of the qualities of "comparability" (lack of bias) and representativeness achieved by sampling in its various aspects, also depend ultimately on randomization (though they may also be governed by other forms of control). Finally, procedures that involve particular methods of sampling or assignment govern not only the proper estimate of error for a valid test of the experimental hypothesis, but also concurrently influence the experiment's precision.

## Validity, bias and sampling

Let us begin by reviewing some distinctions and considerations concerning the validity of experiments and the control of bias. These notions can then be related to the operation of randomization, considered as a key aspect of sampling. Sampling will be considered here with particular reference to the selection and assignment of experimental S's--i.e., students or learners. Later it will also be considered more broadly to include sampling of conditions, tasks and observations or measures, as these affect the representativeness, generalizability or "external validity" of experimental results.

We should first review, so as to have clearly in mind, the basic distinction between the key qualities of validity and sensitiveness as applied to experiments (cf. Chapter 3 p. 42 and Chapter 7 pp. 8 ). Our primary concern will be with validity as an essential condition in any experiment; our secondary concern will be with sensitiveness--particularly precision--as a desirable but less crucial concern (in the long view of scientific history, at least). We want always to try to draw sound conclusions from our data (or else hold our peace or--preferably, and earlier still--stay our hand from useless experimentation). If possible, we then also want to achieve the economy, efficiency and more rapid rate of progress that accrue to an experiment that is conducted in sensitive fashion.

However, despite this distinction in primacy, some of the aspects of sampling with which we shall here be dealing affect both validity and precision in closely interrelated fashion. Therefore, part of the time at least it will be preferable to adopt as the primary frame of reference the operational aspect (e.g., replication or stratification), and then relate this to the resulting effect on first the validity, and second the sensitiveness or precision of the resulting experimental structure.

*Validity*. By an experiment's validity, we recall, we mean basically the attribute which permits us to make an unambiguous interpretation of experimental data with reference to the tenability of the hypothesis which the experiment was set up to test. This means, first of all, achieving absence of bias through appropriate methods of control (including, but not limited to, randomization); and, second, the concomitant ability to derive, directly from the experimental operations performed, a valid estimate of the chance errors to which the experimental measurements are in fact liable. In other words, we must have a valid basis for rejecting the null hypothesis that the treatment variable(s) had no effect (and that, in consequence, only chance effects are present in the data), as well as also being able, because of controls we have employed, to rule out *other* alternate hypotheses.

It may also be useful to review here the distinction made earlier (cf. Chapter 3, p. 41 ) between internal and external validity—a distinction which roughly parallels the one made (Chapter 3, pp. 29-30) between the essential quality of comparability and the also very desirable but in a way less crucial characteristic of representativeness. It will be recalled that *internal* validity—the experimental *sine qua non*— is restricted to the question of whether or not effects observed can be unambiguously attributed to the *specific treatment employed* (i.e., the independent variable *as implemented* in a particular experiment), without regard to whether the results can be generalized to other situations. The question of internal validity is: "*Did* it happen?" (Did these treatments produce an effect, or did these treatments produce a difference in effect, on the subjects' *behavior*.)

External validity, on the other hand, was differentiated as having a meaning roughly synonymous with representativeness or generalizability of results—that is, the extent to which the results of the experiment would furnish a valid basis for predicting what would happen in another situation involving the same hypothesis and the same named independent and dependent variables. Here the question thus might be put as: "Will it happen if we try it again?"

However, the distinction is necessarily one of convenience and is a matter of emphasis, for internal validity can, in fact, be established only in relation to chance errors associated with sampling of observations (always, inter alia, involving sampling of S s--and/or of tasks, and/or of conditions). In other words, since no situation is ever exactly repeated, the concept of chance errors always in some way involves the notion of sampling and hence of inductive generalization; thus, even the matter of so-called "internal" validity must always involve some extrapolation beyond the experiment as it was done. In any case, we want if possible to furnish a valid basis for predicting what would happen in a similar situation to a new sample of similar learners to which the same treatment was applied, or what would happen, if this kind of treatment variation were applied to a new task, at least, for these same learners.

Nevertheless, the distinction is a useful one in that it permits us to make important distinctions in emphasis between factors relating more directly to generalizability ("external validity") and those relating more narrowly to the degree of direct reproducibility ("internal validity") permitted by our data. We might say that an instructional-media experiment's internal validity is supported if, for example, we demonstrate a basis for confidence that the same program, applied under roughly similar conditions to a new random sample of *S* s from the same population, would show similar effects--whereas, with respect to external validity we might raise the further questions of whether the effects of this treatment are applicable also to other kinds of subjects, and/or of whether a similar treatment applied under new conditions or to a new program would show similar effects.

In this sense, external as well as internal validity asks the question of whether, on the basis of experimental results, it is possible to make inferences from the data obtained which bear on the experimental question being asked. From the standpoint of generalizability it is important to note that the appropriate "question" (which, in fact, is really defined operationally by *what was done in experiment*) is *not* necessarily the question that the experimenter had in mind or has stated in his description of the experiment. The grossest instance of this discrepancy is the familiar one where a film is compared with some alternative "treatment" such as a lecture. Here, if the experiment is otherwise satisfactory, one can really only say something about the relative effectiveness of *this particular film* and *this particular lecture*; but the experimenter may nevertheless report (and believe) that he has obtained evidence about a general "method" of instruction.

## Validity and sensitiveness as goals

In making any experimental comparison between two treatments, we want to know whether we are making a *fair*, or sound, or valid comparison, and we also try, when possible, to make an efficient comparison. For the latter purpose we want to impose certain controls so that, even though we have already assured that the comparison is "valid" or "fair" (not systematically biased), those chance (or random) differences between conditions or individuals which may have a bearing on the experiment's sensitiveness will also be minimized. This applies to any suspected or known source of error in an experiment. We may not know how important these errors are, but we must do three things: (1) in the interests of validity, try to eliminate all *systematic* differences; (2) in the interests of sensitiveness (and, thus, of efficiency), try insofar as feasible to control or minimize some of the sources of chance (random) error; and (3) in the interests of both sensitiveness and validity, set up the assignment of individuals, schedules of hours, etc., by a randomizing procedure which will *keep the amount of error under our cognizance* --i.e., permit some valid way of estimating the amount of error that the variability will introduce. The hallmark of a *valid* experiment is *not* the extent to which an experimenter is able to achieve constancy of conditions and keep conditions completely standardized; rather, it is the extent to which, with whatever variability remains, he is able to *estimate properly the errors* that do affect the outcome.

It should be noted that by validity we mean here the extent to which we can be sure that we have not done something that biases the comparison in one direction or another in a systematic fashion, or that leaves errors in the experiment that fail to enter properly into our error *estimate*.

### Accountability of unidentified sources of error

There are always some completely unknown and unsuspected sources of error. Now, obviously, one cannot control or hold constant those things which are unknown. Nonetheless it may often be the case that, even though unknown and unidentified, such sources of error can be taken into account in the error estimates. In order to do so we have to use a randomizing procedure which will assure that these errors are random. How can we do this if we don't know what the sources of error are?

Although we may not know what all the sources of error are, we may be able to establish that they are correlated, or associated, with something we can identify--namely, with the identity of individuals. For example, the process of randomization in assigning subjects to experimental treatment $X_1$ or $X_2$ will enable us to get into our error term any and all unidentified characteristics of verbal ability, brain size, eyelash length, educational level, hair color, etc.--any characteristics that are individual-bound. Some of these factors will be relevant; others won't. But whatever ones are relevant will be taken into account, even though they remain unidentified.

## Systematic and chance differences as sources of error

Sources of error or bias that can make an experiment invalid are, we have noted, those that in some important way make the groups compared differ systematically with respect to something other than the thing we say we are manipulating in the experiment. For example, if we are doing an experiment in which we wish to compare color film versus black-and-white film, there is the assumption that the group which is exposed to the color film does not differ systematically with respect to other factors from the group which is exposed to the black-and-white film. We must always try to insure that relevant population characteristics do not differ systematically for the two experimental treatments, and to do this we must correctly employ randomization at some point in the proceedings. But guaranteeing against systematic differences, or bias, does not mean we have to try to guarantee against chance differences between groups--which, as a matter of fact, one can never really do. And the extent to which we will try to minimize non-systematic differences affects only the sensitiveness of the experiment, not its validity.

However, some of the same sources of error may operate either in systematic fashion or in chance fashion. In particular, it is important to identify two classes of factors, either of which, depending on circumstances, may (a) vitiate (bias) the experimental comparison, or (b) without systematically biasing it in favor of one treatment or the other, still desensitize the experiment and thus make it inefficient. These two classes of factors are: (1) differences in the kinds of people that are assigned to the experimental treatment--individual differences, and (2) differences in the external circumstances--showing conditions of a film, temporal or distribution of effort in using a self-instructional program, for example--that affect the behavior of our subjects. Both of these classes of factors will enter into our subsequent discussion.

## Sampling

Any set of observations on the basis of which effects of an instructional program are to be assessed always represents only a partial sample of the entire "universe" or "population" of instances to which it is hoped the results will apply. A sample is used because it is manifestly unfeasible to make all possible observations under all possible conditions on all possible people.

In speaking of a population from which samples may be drawn we shall be referring here primarily to samples from a population of students or subjects--that is, a population of people. However, it should be understood that the term population (or universe), as a formal term used in the theory of sampling, can refer to any body of observations from which a sample could be drawn. Thus we could speak of a population of subject matters, a population of tests, a population of methods or, as in sampling inspection in quality control, a population of objects from which a sample might be inspected.

The main purpose of drawing a sample is simply to make an estimate of some characteristic of the total population. Examples are estimating the political opinions or sex habits of a country's population through a public opinion poll or "sample survey," or determining the characteristics of a large batch of products in acceptance sampling for industrial quality control. In a well-conducted opinion poll, satisfactory estimates of the responses of the voting public are often made from a sample of three or four thousand interviewees. If we were interested in estimating the average I.Q. of school children in a large city, we could, similarly, use a random sample of 1000 students on which to base a sufficiently accurate estimate of certain characteristics of the total population of perhaps 500,000 students.

In experimental work, on the other hand, we are normally concerned with more than one sample, and our purpose in selecting the samples is to compare groups of individuals who are exposed to different treatments or, in any case, to compare samples of test scores resulting from different instructional treatments. In the former case--estimation of population characteristics from a given sample--the emphasis is on representativeness, whereas in experimental work we tend to stress more the quality of comparability of the samples.

In either case, however, what we are interested in is obtaining an unbiased sample, that is, one which does not differ systematically in its characteristics from the population of which it forms (or is presumed to form) a part, or from another sample with which we wish to compare it. We are also interested in drawing the sample in such a way that the limits of chance error by which it will deviate from another sample or from the total population can be validly estimated.

The simplest way--but not always the most convenient, most feasible, or most precise way--of simultaneously assuring both that a sample is unbiased and that a basis is afforded for making an unbiased estimate of the chance sampling errors is to draw a pure random sample from a defined population. For purposes of estimating the characteristics of a population, for example, we might start out with a complete roster or enumeration of all the members of the population and--using a table of random numbers*--select some satisfactory proportion purely at random. For a large survey such a sample might, for example, comprise 5% of all high school seniors during a given year. The randomly chosen 5% would then afford a basis for estimating the characteristics of that total group of high school seniors.

Stratified or unit sampling. On the other hand, it is often difficult to obtain such a complete roster and it might be very expensive, involving a great deal of travel to locate and test or interview one member of the sample. Consequently, stratified sampling, "area sampling," or sampling of units is frequently employed, the units being chosen at random from a population of units. For example, 25% of all of the high schools in the United States might be chosen at random to include in the sample, with 20% of the students randomly sampled in each high school. This could yield an equally valid estimate, though a precise estimate of the sampling error in this case would differ from that for a purely random sample of the population as a whole.

In addition, aside from matters of convenience, the use of so-called stratified samples is sometimes adopted because under certain circumstances a more precise estimate can be obtained from a given number of "cases" (i.e., from a given size of sample) or, conversely, a more precise estimate can be made with a given sample size. However, using a stratified sample does not increase the validity of the sample, compared with simple random sampling, for representing the population which the sample represents; it only affects the size of the sampling error. Furthermore, a valid estimate of the sampling error must take into account the kind of sampling that actually was done--though, unfortunately, it is not uncommon for the statistics used in estimating sampling error, when unit sampling or stratified samples have in fact been used, to be calculated as if the sample were really a purely random sample of the entire population. We shall return to this consideration presently as it applies in the special use of unit sampling to select individuals for assignment to experimental treatments.

Randomization in sampling. In any sampling a valid estimate of the sampling error depends on some form of randomization--either across-the-board randomization, or randomization of the units that are chosen, and/or the individuals that are selected within any unit or stratification category. As we have previously stressed, this applies in particular in the selection of samples of individuals to be assigned to experimental treatments (as it does, also, in the selection of samples for survey purposes in estimating population characteristics).

An important aspect of experimental design is thus to avoid selection bias, which may potentially result from the assignment of a group of subjects to receive an experimental treatment on some basis other than unbiased sampling involving proper randomization. A biased sample will differ initially (that is, without reference to the effect of the experimental treatments) from groups with which it is to be compared and thus invalidate the experimental comparison. Proper randomization will always preclude this source of systematic bias--a special form of which is the so-called "self-selection bias" discussed earlier (Chapter 3 , p. 20 ff) in which those exposed to an instructional program consist of those who voluntarily choose to do so. Another special source of selection bias which will be discussed later (Chapter 15) underlies the spurious "effects" that can appear as a consequence of regression.

It should be remembered that randomization has two related purposes. The first is to remove the possibility of obtaining a systematically biased sample; the second is to generate a valid estimate of sampling error. It is very important to remember that the first of these may be achieved while failing to achieve the second. The proper thing is always to achieve both jointly. This is done if and only if the number of independent acts of randomization in the sampling or assignment of experimental units (subjects and/or tasks and/or conditions) to treatments is the basis for the number of degrees of freedom in the error estimate (see Chapter 8, p. 48 ). If we were to toss a coin to determine once and for all which one of just two available classrooms will receive one treatment (e.g., color film) and which will receive the alternative treatment (e.g., black-and-white film), we have achieved lack of bias, but we do not have a proper basis for estimating experimental error. The reader should once again be reminded that establishment of comparability of samples in a true or rigorous experiment must flow directly from the sampling operations that were performed, and not from some *post hoc* statistical justification.

The necessary randomization process need not, however, be *simple* randomization. For example, let us say we wish to sample a group of people, and to take account of the fact that people differ with respect to I.Q. In this case, we can either draw two samples out of this group completely at random, or we can rank them according to I.Q. and take the adjacent pairs from which, on the basis of coin-tossing or referring to random numbers, we assign *one subject from each of the pairs* to experimental treatment I and the remaining subject of the pair to experimental treatment II. What we determine randomly here is merely *which* individual of the two in each pair goes to which experimental treatment. In doing that we have not two undifferentiated random populations, but two sets of pairs; and we estimate our error accordingly, as in the examples in Chapters 4 and 8. If I.Q. is highly correlated with the criterion variable, this will produce a more sensitive experiment than if we merely assigned the subjects *completely* at random.

Once again we emphasize that such a pairing or "matching" procedure does not increase the validity of the experiment; it simply increases the precision and therefore the "sensitiveness" of the experiment. That is, it *may* increase the precision (though not necessarily--see Chapter 11); but in any case we can *not* improve on the validity of an experiment which employs straight randomization. Furthermore, let us repeat that simple pairing, as such (that is, without randomizing the assignment of members of pairs to treatments), or other forms of matching or stratifying are not a substitute for randomization. There is no way that one can ever get a valid estimate of experimental error except by randomizing at some point in the experiment. This is not to say that one might not happen to arrive at a true conclusion, but there is no way of arriving at a valid estimate of the error term unless we randomize, either across the board, within pairs or strata, or by units--and do so in such a way that *the number of acts of randomizing govern the number of degrees of freedom upon which our error estimate is based.*

### *Designs using pairing of subjects*

In the experiment described in the second part of Chapter 4, and the similar experimental variation yielding quantitative scores for analysis described in Chapter 8, *pairs* of subjects were used--each member of a pair being assigned at random to a treatment. (If we had had three or four treatments instead of only two, we could similarly have made random assignments among triplets or quadruplets of subjects instead of pairs.) Our pairs were chosen, furthermore, as pairs of individuals who were "matched" on some relevant characteristic such as I.Q., learning ability, or scholastic achievement. We might also have used *two or more* such matching characteristics in combination as the basis for pairing--that is, for identifying which particular individuals constituted each pair.

The objective of the pairing was to increase the precision of the experiment by reducing the amount of uncontrolled sampling variation, while still maintaining the essential operation of randomizing to afford an unbiased comparison and a valid basis for estimating the error that was not taken into account by the pairing process. (Remember that we can never remove all of the error, by pairing or any other control procedure; often we cannot even remove very much of it this way.)

*Pairing and stratification.* As will be further discussed later (see

Chapter 11, p.        ), pairing is the extreme form of the method of stratified sampling. In less extreme forms, we merely group together a number of subjects on the basis of some identifiable characteristic and then assign randomly within such strata. This method may be used because it is the only feasible method available to us--as when we randomize within the "stratifications" determined by factors outside our control, such as prior assignment to classroom groups. It may also be a method which we choose to use, even when not forced to do so, because randomization within strata, rather than across the board, can reduce sampling error. Here, for example, we might assign Ss at random within the categories of A, B, C, and D students, or within groups of advanced, regular, and below average students. When the stratifications are so small or fine that they consist of single pairs (or triplets, etc.), then we are, of course, using the method of pairing. In this case, to the extent that the basis for pairing is the most relevant one, we are removing all of the error that can be removed by stratified sampling. We then merely take the differences between scores of pairs of subjects as our basic observation, as in the examples in Chapters 4 and 8.

For three or more treatments, however, such a computation of paired differences becomes more complex. For example, among the trio of treatments $X_1$, $X_2$, and $X_3$ we would have three sets of differences, namely $X_1 - X_2$, $X_2 - X_3$, and $X_1 - X_3$, to deal with. However, the basic logic that would be involved in the analysis of variance which we would probably perform in such a case is essentially the same as that underlying the analysis of paired differences in the two-treatment paired-scores analysis. (It might be noted that in the analysis of variance, in such a case, the correlational source of variation is called "interaction variance.")

## Simple Randomization" Design

Of all the forms of experimental structure (methods of sampling in the [illegible] to be assignment of subjects to treatments), the simplest in conception, *purely* random assignment of subjects. If this method is scrupulously employed, it is perhaps also the safest of sampling methods, since it involves the fewest complications. As has been noted, however, it is *not* necessarily the most precise method (in the sense of producing the smallest amount of sampling error), although under some conditions it may be no less precise than other, more involved methods. Improvement of the precision of pure random sampling by various methods of pairing or stratification will, in fact, depend on the availability of a relevant basis for such pairing or stratifying in terms of variables which are appreciably correlated with the criterion performance to be achieved by the instruction (see Chapter 11, p. ).

However, despite its conceptual simplicity, we did not take up purely random assignment as our initial example in Chapters 4 and 5 because the statistical tests that apply in the case of purely random sampling are, while relatively simple, nevertheless somewhat less simple or "transparent" than in the case of paired measures. This possibly paradoxical-seeming fact holds true both in the case of qualitative comparisons in which only the superiority of one treatment is taken into account, without reference to measurement of *amount* of superiority (as in the case of the second experiment in Chapter 4) and also in the case where quantitative variables are used (as in the use of the *t*-test for paired measures described in Chapter 8). Accordingly, we chose the paired-measure situation for our initial experimental examples, in order to utilize statistical tests in which the logical steps could be

---

***Footnote to page 10-16**

**We could accomplish the same result less neatly by working with the mean scores of the pair members within each experimental treatment, rather than using the paired differences and then using the formula given on page 9-29 (in which the resulting error reduction due to pairing is taken into account through the correlational terms); this would also serve to remove from our error estimate that amount of error which has in fact been taken account of by the pairing process.**

correlational source of variation is called "interaction variance.")

## "Simple Randomization" Design

Of all the forms of experimental structure (methods of sampling in the assignment of subjects to treatments), the simplest in conception may be considered to be purely random assignment of subjects. If this method is scrupulously employed, it is perhaps also the safest of sampling methods, since it involves the fewest complications. As has been noted, however, it is not necessarily the most precise method (in the sense of producing the smallest amount of sampling error), although under some conditions it may be no less precise than other, more involved methods. Improvement of the precision of pure random sampling by various methods of pairing or stratification will, in fact, depend on the availability of a relevant basis for such pairing or stratifying in terms of variables which are appreciably correlated with the criterion performance to be achieved by the instruction (see Chapter 11, p. ).

However, despite its conceptual simplicity, we did not take up purely random assignment as our initial example in Chapters 4 and 8 because the statistical tests that apply in the case of purely random sampling are, while relatively simple, nevertheless somewhat less simple or "transparent" than in the case of paired measures. This possibly paradoxical-seeming fact holds true both in the case of qualitative comparisons in which only the superiority of one treatment is taken into account, without reference to measurement of amount of superiority (as in the case of the second experiment in Chapter 4) and also in the case where quantitative variables are used (as in the use of the t-test for paired measures described in Chapter 8). Accordingly, we chose the paired-measure situation for our initial experimental examples, in order to utilize statistical tests in which the logical steps could be followed as directly and completely as possible. We now turn to the case of statistical treatment for unpaired measures--that is, for experiments in which purely random (or, "simple random") sampling is used as the basis for assigning subjects to experimental treatments, with no pairing or other stratification.

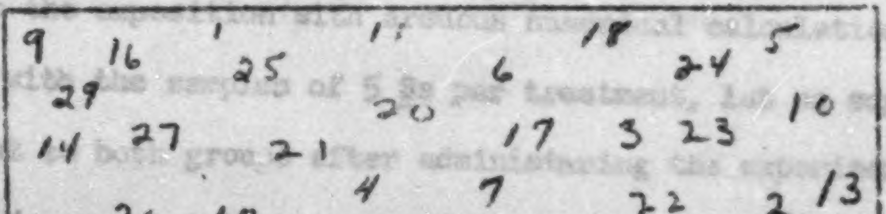
In the rest of this discussion we shall for convenience discuss the comparison of experimental treatments in terms of the differences in means of sets of scores, with the understanding that a similar logic (through different specific probability calculations) would apply when we use other measures such as percentages, medians, or simply the number of students obtaining the scores above or below some specified value. Also, the logic of simple random-sampling comparisons would apply even if (as in our "spelling bee" experiment in Chapter 4) we only made dichotomous distinctions such as those between pass and fail instead of employing quantified measures of some sort for each subject. For unpaired measures with such qualitative data the so-called "median test" would apply; using ranked data, the Wilcoxon test for unpaired replicates could be used if there were equal numbers of subjects in both of two experimental groups, or the Mann-Whitney test if there were not an equal number of subjects in experimental groups

Use of simple random sampling. One of several reasons why we may choose simple random sampling in some instructional-media experiments is that we may have no adequate basis for pairing our subjects. Another consideration is that it may be necessary or desirable to use unequal numbers of subjects in two experimental groups, so that it is impossible to pair them up one-for-one across treatments. Use of purely random sampling as the basis of assignment of subjects to treatments can still provide us with an entirely valid basis for estimating error, regardless of whether an equal number of students are assigned to the several experimental treatments. In such cases the chance error--that is, variation due to all differences among subjects except those caused by experimental variable--should enter into the randomization process and thus into our estimate of error. This contrasts with having some of it removed (both from the actual error in the experiment and from the estimate of that error) by pairing or stratifying.

In a sample-randomization experiment we simply assign subjects on a purely random basis to two treatments and compare the differences in scores between the two treatment groups with the amount of difference to be expected by chance on the basis of purely random sampling. The latter is determined by the number of subjects in each sample and the amount of variability that is characteristic of their performances. The null hypothesis now implies that the two sets of scores obtained for the subjects in the two treatments are two random samples from the same population of scores. (On the other hand, in the previous case of directly comparing the differences between the members of pairs of subjects, the appropriate null hypothesis was that these sets of paired differences in scores were a random sample from a population of difference scores whose true mean was zero.)

## An example of use of a simple randomized design

Let us use the following representation to indicate the assignment, purely at random, of a sample of subjects to two experimental groups $G_1$ and $G_2$, which will receive two treatments $X_1$ and $X_2$. We will pick 5 Ss at random for each treatment. The population of available Ss is composed of, say, 30 individuals, located in no particular order in a room, on a playfield, etc., like this:



However, though the order looks haphazard enough, we don't know that it is really random--quite possibly all of the students at the left are boys, those at the right girls, or those at the top are good students and those at the bottom poor students. So, making explicit use of a table of random numbers we chose, from the 30, two samples of five each, which turn out to be students No. 3, No. 17, No. 24, No. 20 and No. 5, to be assigned to treatment $X_1$, and students No. 28, No. 21, No. 4, No. 19 and No. 16 as the sample to be assigned to treatment $X_2$.

There is, of course, nothing to preclude our using all of the available subjects and dividing them randomly into the two treatment groups and, indeed, this is a procedure which would ordinarily be followed. It is, also, unnecessary that the two groups be equal in size, though it is convenient and customary to work with equal-sized groups, which generally provides the most efficient way of utilizing a given number of available Ss (and also somewhat simplifies the computations needed for the appropriate statistical test). ¶ However, use of two small and equal-sized samples from an available pool of Ss is sometimes quite sufficient for E's purposes, and may be well advised, especially if the treatments are such as to require extensive individual work with subjects, or if the scoring procedure for each subject involves considerable work. In any case, the use of a small number of cases here will be convenient to reveal the important principles in comparing randomly assigned groups without cluttering up the exposition with arduous numerical calculations.

Working with the samples of 5 Ss per treatment, let us suppose that we give a test to both groups after administering the experimental treatments. We now want to test the hypothesis that $X_1$ is better than $X_2$, and to do so must rule out the null hypothesis that the obtained

* Footnote to page 10-20

If students were being tested sequentially one at a time, we might simply determine at random which Ss, in order of appearance, would get which treatment.

difference in the distribution of scores could readily arise just as a consequence of random sampling errors. As has been noted in Chapter 8 (pp. 8-39, 8-40) Obviously the rejection of the null hypothesis will be affected both by the size of the difference between the means and by the variability of the measures within each sample. Let us review the implications of this fact. For example, suppose the mean for $X_1$ is 15.0 and that for $X_2$ is 10.0. The distribution of scores within each group might have very little variability; for example it might be as follows:

```
 X1                                        x
                                           x
                                      x    x    x
 score:  9    10   11   12   13   14   15   16   17
              x
              x
 X2      x    x    x
```

The two distributions here are far separated, without overlap and the values for each group are highly consistent--i.e., have very small variance. Thus we might easily make the correct prediction that an appropriate statistical test would disprove the null hypothesis; it seems evident that two such contrasting distributions would not be likely to arise merely from chance.

On the other hand, with the same means the two samples might look like this:

```
                         x    x         x         x    x
 7    8    9    10   11   12   13   14   15   16   17   18
 x    x         x         x    x
```

Here, with more variability in relation to the difference in means, so that two distributions show partial overlap, it is less clear that the difference in the means of 10 and 15 for the two samples might not arise as a chance matter.* (cf. similar example on p. 40 of Chapter 8).

However, with the *same* amount of variability within samples, as in the second case, the null hypothesis would *not* appear reasonable if the difference between the means were, say, 15 instead of only five, giving two non-overlapping distributions like this:

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | x | x | | x | | x | x |
| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| x | x | | x | | x | x | | | | | | | | | | | | | | | |

The question now is how to determine, for any given difference in means and any given amount of empirically observed variability, the probability that such a pattern of results would arise by chance sampling. This we have already done (Chapter 8, p. 55 ff.) for the case of paired measures where, in effect, the pairing enabled us to reduce the two sample distributions to a *single* distribution of paired differences and to determine the probability that the obtained mean difference would occur as a chance deviation from a true mean difference of zero.

In that case, when the population variance was unknown, using the *t* distribution we calculated an estimate of the sampling variation of means based on the size of the sample and the variance of the obtained distribution of differences. We first employed the relation $\tilde{\sigma}^2 = \frac{1}{N-1}\sum(Y-\bar{Y})^2$

(or, using scores y, each calculated as a deviation from the mean,
$\tilde{\sigma}^2 = \frac{\sum y^2}{N-1}$ ) to obtain an estimate of the population variance. Then, using the algebraic fact that the variance of means of samples

of size N is $\sigma_M^2 = \frac{\bar{\sigma}^2}{N}$ or, as estimated from a sample, $\frac{\tilde{\sigma}^2}{N}$, we calculated $t = \frac{\bar{Y}}{\sigma_M}$.

Now, however, we are dealing not with a single distribution of differences between pairs, but with two sample distributions, one for each treatment. We therefore have to inquire what is the variance of a distribution of unpaired, and thus uncorrelated, differences. It can be shown by very simple algebra that the sampling variance of uncorrelated differences, $Y_1 - Y_2$ is $\sigma_1^2 + \sigma_2^2$, where $\sigma_1^2$ is the variance of the $Y_1$'s and $\sigma_2^2$ is the variance of the $Y_2$'s. This is true for differences between sample means as well as between single measures; that is

$$\sigma^2_{M_1 - M_2} = \sigma^2_{M_1} + \sigma^2_{M_2}.$$

For unpaired samples we could then calculate readily

$$t = \frac{M_1 - M_2}{\sqrt{\sigma^2_{M_1 - M_2}}} = \frac{M_1 - M_2}{\sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}}}$$

if we knew the true population variance $\bar{\sigma}^2$, since

$$\sigma^2_{M_1} = \frac{\bar{\sigma}^2}{N_1} \text{ and } \sigma^2_{M_2} = \frac{\bar{\sigma}^2}{N_2}.$$

When we do not know $\tilde{\sigma}^2$, however, we must estimate it. But in the present case we have two samples on the basis of which $\tilde{\sigma}^2$ can be estimated. In this case, it can be shown that the difference $M_1 - M_2$ can be evaluated in terms of the known distribution of *t* if, on the null hypothesis, we make an unbiased estimate of $\sigma^2$ in terms of the variability of the observations in each sample from their own sample mean. This is done by estimating $\tilde{\sigma}^2$ as

$$\frac{\Sigma(Y_1 - M_1)^2 + \Sigma(Y_2 - M_2)^2}{N_1 + N_2 - 2}.$$

The denominator $N_1 + N_2 - 2$, or $(N_1 - 1) + (N_2 - 1)$ by which the sum of squares of deviations from the sample means from the combined samples is divided to estimate the population variance again represents the number of independent deviations used in forming this estimate—since, just as in the "one-sample" case with paired differences in Chapter 8—one degree of freedom is lost in the computation of deviations from each mean.

Using $y_1$ and $y_2$ to stand for deviations of the observed scores from their respective sample means, $M_1$ and $M_2$,

$$\tilde{\sigma}^2 = \frac{\Sigma y_1^2 + \Sigma y_2^2}{N_1 + N_2 - 2}.$$

Since the variance of a mean (see Chapter 8, p. 47 ) is $\frac{\tilde{\sigma}^2}{N}$, and the desired sampling variance of the difference between means $I_1 - I_2$ is therefore $\sigma^2_{M_1 - M_2} = \sigma^2_{M_1} + \sigma^2_{M_2} = \frac{\tilde{\sigma}^2}{N_1} + \frac{\tilde{\sigma}^2}{N_2} = \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\tilde{\sigma}^2$

$$= \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\Sigma y_1^2 + \Sigma y_2^2}{N_1 + N_2 - 2}\right)$$

If the sample sizes are equal $(N_1 = N_2)$, this is simply

$$\sigma^2_{M_1 - M_2} = \frac{2(\Sigma y_1^2 + \Sigma y_2^2)}{N(2N-2)} = \frac{\Sigma y_1^2 + \Sigma y_2^2}{N(N-1)}.$$

In any event (regardless of whether the N's are equal) we can then calculate $\underline{t} = \frac{M_1 - M_2}{\sqrt{\sigma_{M_1 - M_2}}}$, and compare the obtained value of $\underline{t}$ with the $\underline{t}$ distribution for $N_1 + N_2 - 2$ degrees of freedom.

Let us now apply this to a simple example, for which we may assume that the scores obtained by the five $X_1$ subjects and the five $X_2$ subjects happen to be exactly the same as those used in the example in Chapter 8 for paired scores. Since the scores are however now for unpaired random samples, we will use the relationships developed just above instead of those for distributions of paired differences.

The scores for $X_1$ and $X_2$ are, let us then suppose, as follows:

| Group $X_1$ | | Group $X_2$ | |
|---|---|---|---|
| Subject | Scores ($Y_1$) | Subject | Scores ($Y_2$) |
| No. 3 | 19 | No. 28 | 9 |
| No. 17 | 10 | No. 21 | 2 |
| No. 24 | 13 | No. 4 | 8 |
| No. 20 | 18 | No. 19 | 15 |
| No. 5 | 15 | No. 16 | 16 |
| | Sum 75 | | Sum 50 |
| Mean | 15.0 | Mean | 10.0 |

It may be convenient, though not necessary, to arrange the scores in order of size (with larger samples, where several individuals may get the same score use of the resulting frequency distributions will simplify computation). In doing so, we may also dispense with the subjects' identifying numbers which have no relevance to our purpose. We may also note that such rearrangement of scores according to size would be wholly illegitimate if we were then to consider corresponding scores as pairs, and use paired differenses. Any such pairing must always be done on some prior basis rather than in terms of the scores themselves if the analysis is not to be wholly invalidated.

We may also conveniently indicate the deviation of each score from its group mean as follows:

| Group $X_1$ | | | Group $X_2$ | | |
|---|---|---|---|---|---|
| | Scores $X_1$ | Deviation scores $y_1(X_1 - M_1)$ | | Scores $X_2$ | Deviation scores $y_2(X_2 - M_2)$ |
| | 19 | 4 | | 16 | 6 |
| | 18 | 3 | | 15 | 5 |
| | 15 | 0 | | 9 | -1 |
| | 13 | -2 | | 8 | -2 |
| | 10 | -5 | | 2 | -8 |
| Sum | 75 | 0 | Sum | 50 | 0 |
| Mean | 15.0 | 0.0 | Mean | 10.0 | 0.0 |

The deviation scores in each case sum to zero, as they must. In this simple example they are easy to calculate (we chose the example so they would be), since the means come out even and hence the deviation scores are all integers. Thus we can readily calculate the required variance estimates directly from these deviations as $\frac{\sum y_1^2 + \sum y_2^2}{N_1 + N_2 - 2}$.

With most actual data, often involving larger N's and seldom yielding such convenient mean values, we might wish to use a different procedure, especially if we have a calculating machine handy, and work with the squares of original scores instead of first computing the deviation scores, and using a more complex but equivalent formula that would obviate calculating the deviation scores as such.** Doing so tends to obscure the logic somewhat and thus is avoided for our present purpose. $\sum y_1^2$, for our example is $(4)^2 + (3)^2 + (0)^2 + (-2)^2 + (-5)^2$, or 16 + 9 + 4 + 25, or 54 (which, of course, we could also have calculated by obtaining the deviations without first rearranging the data). Similarly,

$\Sigma y_2^{\,2}$ is 36 + 25 + 1 + 4 + 64, or 130.

Thus

$$\tilde{\sigma}^2 = \frac{\Sigma y1^2 + \Sigma y2^2}{N_1 + N_2 - 2} = \frac{54 + 130}{5 + 5 - 2} = \frac{184}{8} = 23.$$

Now to assess the reliability or "significance" of the difference in relation to expected chance sampling differences between two sample means each of size N = 5, we calculate

$$\sigma^2_{M_1 - M_2} = (1/5 + 1/5) = 2/5(23) = 9.2,$$

and $\sigma_{M_1 - M_2} = \sqrt{9.2}$

whence $$t = \frac{15.0 - 10.0}{\sqrt{9.2}} = \frac{5.0}{\sqrt{9.2}} = \frac{5}{3.04} = 1.65.$$

The table of the distribution or *t* shows for eight d.f. (5 + 5 - 2) shows the probability of getting a value of *t* of 1.65 in a specified direction is about .08, which would not of course quite meet the more stringent usual criteria for rejecting the null hypothesis. [Comparison of this value with that obtained using the same scores as *paired* measures (Chapter 8, p. 56) indicates that the pairing was, in this case, worthwhile even though it reduced the d.f. from eight, for unpaired scores, to four for the paired scores. Whether such an advantage of pairing would be realized depends, as noted earlier, on the size of the correlation between the paired scores: see further discussion in Chapter 11.]

Similar calculations apply in any experiment using simple random assignment of *S*s to two treatment groups: estimate the population variance as $\Sigma y_1^{\,2} + \Sigma y_2^{\,2}$ divided by $N_1 + N_2 - 2$; calculate the variance of sample means (i.e., the square of the standard error of each mean) by dividing by N; add the two mean variances, and divide the square root of their sum into the mean difference to get *t*. The calculation is thus only slightly less simple than that for paired measures described in

Chapter 8 (pp. 54-56). In Chapter 11 we shall consider the case where we have randomized subjects within each of several "stratified" subgroups. But first we must consider another extremely important situation often encountered in instructional-media experiments involving class groups--namely, the situation where entire pre-formed groups, rather than individuals, are assigned as units to experimental treatments.

## Experiments Employing a Number of Groups of Subjects

There are two important ways in which assignment of subjects to two experimental groups in an instructional experiment can, while involving randomization that permits a valid estimate of error, depart from purely random sampling. It is of the utmost importance to differentiate these clearly, because (1) they are frequently confused, and (2) they are diametrically opposed with respect to the basis on which a valid estimate of experimental error needs to be arrived at. The basis for confusion arises in part from the fact that a diagrammatic or tabular representation of structural pattern looks the same for both of the two methods, even though they represent entirely different experimental structures. The crucial difference in structural pattern and resulting basis for estimating error lies in the fundamental question of how--i.e., by what acts of randomization--Ss are assigned to different "cells" or boxes in the pattern.

In descriptions of experimental "design", (in the restricted sense we have called experimental "structure", ) it is not uncommon to draw a small box-like chart or table to represent the experimental groups. Sometimes the number of "cases" or "N"--by which for the present we will mean the number of Ss assigned to each treatment--are placed inside the boxes. Usually this is done only where the design gets somewhat complex, but we can do it for the simplest of designs, say an experiment comparing two treatments, $X_1$ and $X_2$, thus:

| $X_1$ | $X_2$ |
|---|---|
| 8 | 8 |

Here we have eight subjects for $X_1$ and eight for $X_2$. A similar diagram might represent three treatments; for variety, let us put ten Ss in the box or "cell" for each treatment, thus:

| $X_1$ | $X_2$ | $X_3$ |
|---|---|---|
| 10 | 10 | 10 |

Total Ss = 30.

A slightly more complex design (structure) appears in the form of a so-called two-by-two table, or fourfold table, representing the combination of two variables, X and Z, each with two conditions. For the sake of specificity, we might consider $X_1$ and $X_2$ to be two forms of repetition used in a program, and $Z_1$ and $Z_2$ to be two ways of prompting the students' responses. Regardless of the specific variables, we can draw a structure diagram like this:

|  | $X_1$ | $X_2$ |  |
|---|---|---|---|
| $Z_1$ | 8 | 8 | 16 |
| $Z_2$ | 8 | 8 | 16 |
|  | 16 | 16 | 32 |

Here we have 16 subjects apiece for $X_1$ and $X_2$ and 16 subjects apiece for $Z_1$ and $Z_2$. But we have also combined the treatment so that:

8 Ss get the combination $X_1$ and $Z_1$

8 Ss get the combination $X_2$ and $Z_1$

8 Ss get the combination $X_1$ and $Z_2$

8 Ss get the combination $X_2$ and $Z_2$

This particular "2 x 2" structure (as well as elaborations of it, such as 2 x 3, 3 x 3, or 2 x 2 x 2, or 3 x 3 x 2) is one that is quite commonly used in experiments, and we shall return to it later. [Any such combination of two or more variables in such a way as to have a "cell" with Ss in each category for one variable within each category of one or more other variables is called a "factorial" design or structure.] The 2 x 2 structure is the simplest factorial structure we can have for an experiment.

Now let us look at two such structure diagrams that each involve eight cells, which we can draw in a "2 x 4" pattern. In each the horizontal dimension represents two treatment conditions, $X_1$ and $X_2$. The categories of the vertical dimension we will simply call a, b, c, and d, not worrying for the moment what variable these four categories might represent.

Here are the two "designs," for each of which we have indicated, as an example, 6 Ss to each cell.

Design I

|  | $X_1$ | $X_2$ |
|---|---|---|
| a | 6 | 6 |
| b | 6 | 6 |
| c | 6 | 6 |
| d | 6 | 6 |

Design II

|  | $X_1$ | $X_2$ |
|---|---|---|
| a | 6 | 6 |
| b | 6 | 6 |
| c | 6 | 6 |
| d | 6 | 6 |

We said two designs; but they look identical. There are two designs; they are not identical--they require entirely different forms of analysis. But the difference cannot be told from the diagram or the fact that each is described as a "2 x 4 design with 6 Ss per cell." There is no way to know what the difference is except to find out what was actually done in the experiment.

By "done in the experiment" we have in mind here, however, only one particular aspect of what was done--namely, the sampling or assignment of Ss to the cells. (We need not, for our present purpose, ask what is $X_1$ or $X_2$, whether we used before and after paired measures, what the dependent variable is or how it was measured, etc., etc., etc.) The basic difference hidden behind the two like-looking "designs" is a question of how the subjects were assigned to the cells. To bring out this difference clearly in a special but very important respect, let us now consider that a, b, c and d represent some sort of existing groups (or pairs of groups) of *Ss*--classrooms, schools, or districts, for example--from whom our subjects are to be chosen for assignment to two treatments, $X_1$ and $X_2$.

"*Design I*." In one of the designs--call it Design I--we might use four classrooms of students--call the classes a, b, c and d. *Within* each class we randomly assign half the students to one of the two treatments, $X_1$, and the rest to $X_2$. For convenience let us assume small, equal-sized classes of 12 students each (the actual size is immaterial; there might be 12, 20, or 34, or any other even number; **if there is an odd number, we could for simplicity discard the data for one student at random). The** essential point is that the *Ss* were randomly divided, or allocated, to treatments randomly *within each* class. With 12 students per class it is easy to see that the one of the two diagrams at the top of page 10-31 fits what we have done. We have arbitrarily chosen to think of the one at the left, Design I, for the present pattern of assignment. The trouble is that an identical diagram would also fit the quite different assignment operation which we shall now describe.

"*Design II*." The circumstances are now different. Instead of four classes with 12 students, each divided randomly into two groups of 6 *Ss* apiece, let us now imagine that we have *eight* small classes, with only six *Ss* in each, to begin with. And instead of assigning *individual* *Ss* independently to treatments, we are in this second case going to assign *a whole class at a time* to treatments--four classes to each treatment. In doing so we will again wind up with six *Ss* in each box of a "2 x 4" design diagram--say, the right hand one of the two on page 10-31, which we have already labeled "Design II." The two diagrams--Design I and Design II--do properly look alike: each has six *S*s in each cell. But the way the *S*s *got there* is entirely different--and so are the patterns of experimental error to which the two designs are liable.

The "2 x 4" label thus can be most seriously misleading unless it is indissolutly coupled with a clear statement of *what was done in the experiment*--specifically, what was done in assigning the subjects to the classes. For it is evident that a valid estimate of the error is quite different for the two designs, and that in order to base it on, in Fisher's terms (1960, Sec. 19): "...on the effects of the very same causes of variation as have produced the real errors in our experiment," we must particularly note the way in which *S*'s were assigned to treatments. Let us here reiterate that an aspect of this assignment that is crucial in determining the proper estimate of experimental error is the *number of independent acts of randomization* performed in assigning the *S*'s.

### *Individual and group sampling in experiments*

Many experiments on instructional programs (e.g., films) are conducted in real classroom situations. This means that often the experimental treatments are assigned to already-formed *class groups* (as in Design II) rather than involving randomization among *individuals* (as in Design I). This poses special requirements for obtaining a valid measure of experimental error. If these requirements are not met, there is no formal way to assure that the groups compared are "comparable"--that is, in that the comparison is not biased by factors other than the experimental variable which may introduce a systematic though perhaps unidentified bias in favor of one group or another. There are really two aspects to the problem. One is *selection bias* introduced by correlation between the basis for grouping and the dependent variable; the other is the non-independence of selection which violates the definition of randomness for assessing experimental error.

One course of action frequently adopted through ignorance or through *seeming* necessity is to depart from the model of a true experiment (i.e., from the condition that the operations of randomization as actually performed generate directly the estimate of chance error) and settle for a "quasi-experiment" (see p. 95 ff. below) in which *E* then tries to caulk the holes in his design as well as possible by patching-up techniques of matching or equating. For the present let us stick to true experiments in which there is a valid basis for estimating experimental error, and thus consider the requirements necessary to this end--that is to say, as Fisher (1960, Sec. 20, pp. 41-42) states:

> "To guarantee that... *the very same causes that produce our real error shall also contribute the materials for computing an estimate of it*.

As Fisher further notes:

> "The logical necessity of this requirement is readily apparent, for, if causes of variation which do not influence our real error are allowed to affect our estimate of it, or equally, if causes of variation affect the real error in such a way as to make no contribution to our estimate, this estimate will be vitiated, and will be incapable of providing a correct statement as to the frequency with which our real error will exceed any assigned quantity; and such a statement of frequency is the sole purpose for which the estimate is of any use. Nevertheless, though its logical necessity is easily apprehended, the question of the validity of the estimates of error used in tests of significance was for long ignored,

> and is still often overlooked in practice. One reason for this is that standardised methods of statistical analysis have been taken over ready-made from a mathematical theory, into which questions of experimental detail do not explicitly enter.

### *Valid analysis with individual or group sampling.*

The two designs outlined above exemplify the two general ways in which validity may be attained in a group experiment while adhering to the requirements for a true experiment. The first is by sampling (assignment to treatments) of individual subjects rather than groups. The feasibility of this depends on the kind of treatment and conditions involved (see below). The second, where assignment to treatments must be made by intact, pre-formed groups (e.g., classrooms) rather than by individual, is to use the group rather than the individual as the unit of statistical analysis in obtaining estimates of experimental error (cf. Hovland, Lumsdaine, and Sheffield, 1949, p. 327; and Lindquist, 1953, Chapter 7 for examples and further discussion). Only in this way can one fully heed Fisher's warning (1960, Sec. 17) that, "Whether these estimates are valid, for the purpose for which we intend them, depends on *what has been actually done*."

Notice that in "Design I" above there were *12* "independent acts of randomization" *within* each class, or a total of *48*. (Actually they are not quite independent if we start with only 12 *S*'s per class, since once the quota of six for each class is filled up the remaining assignments are automatically determined if we limit each treatment group to six. It would be a "purer" case if we were to assign six to each treatment from a large pool representing a class of indefinite size, or at least one of more than 12 students. In any case, however, all 12 are randomly assigned in the sense that at the outset each of the *S* s in this group has an equal chance of getting into either group, and which group he gets into is determined by chance on an individual basis, not for the class group as a whole.)

In "Design II," by contrast, we make a total of *only eight* random assignments, four classes to each group. (Again it would be purer mathematically, though generally less realistic in terms of the way instructional experiments are conducted, if we draw four classes for $X_1$ and four for $X_2$ from a large pool of available classes.)

What was randomly assigned in Design I was *individual subjects*; what was randomly assigned in Design II was *intact class groups*. The psychometric tradition strongly suggests the tendency to always consider "N" (the number of cases in a statistical analysis) to mean the number of people, but this is true only in Design I; in Design II, the proper "N" is the number of *classes*, because it is the class unit that is assigned randomly to treatments and that therefore represents an *independent* observation--one independent observation *per class*.

In Chapter 11 we shall return to the analysis of data for an experiment conducted in accordance with Design I--a so-called randomized block design in which individual *S* s are randomized within each of several classes, "blocks," "stratifications," or "replications" (for the present at least, we can regard all of these as amounting to the same thing). Here we want to consider further the general class of designs that involve *unit sampling*--the sampling or assignment *as a unit* of any kind of pre-formed or pre-established or pre-selected groups of people. These might be classroom groups, or schools, districts or cities. (For some kinds of problems only whole communities or cities would do, not because we require *large* sampling units but because we need *isolated* units. Thus random assignment of two kinds of *broadcast* TV or radio programs for a rigorous experiment would probably require random assignment of the two treatments to a number of small local stations scattered over a considerable geographic area.)

## Importance of the unit-sampling question in instructional-media experiments

The problem of *unit sampling* or *group sampling* (or assignment) is given special attention here for several reasons. For one thing, it is frequently used *de facto* but with the data improperly analyzed on an "as if" basis--that is, *as if* the *Ss* had been assigned randomly as individuals when in fact they were assigned *en bloc* a whole class at a time.

For another thing, the experimenter on instructional methods in schools generally finds the class as the natural, pre-established unit. The tendency to treat the group or class as the administrative unit in experiments is perhaps, if anything, even stronger when we are dealing with films, TV or other "group media" of instruction--although, as further discussed elsewhere, the use of media which are normally group versus those which are individually paced has no necessary implications for the use in an experiment of (a) group versus individual *assignment*, of (b) group versus individual *instruction* in the experimental-treatment situation, or of (c) *testing* the effects in groups or individually; and, furthermore, the decisions on (a), (b) and (c) can also, in principle at least, all be made independently.

A third reason for giving special attention to the unit-sampling situation is that, whereas often E has no choice and *must* assign treatments on a unit or class basis, on other occasions the determination of whether to use unit or individual assignment *is* an available choice in designing an instructional-media experiment; E should therefore understand the advantages and disadvantages of either choice in terms of their consequences for the requirements of valid analysis as well as their implications for experimental precision. He should also be aware of some other factors that may condition the design decision, and of practical ways in which the individual-assignment choice may be implemented, even though the initial basis for sampling in the choice of groups of Ss to work with, in an instructional-media experiment, may necessarily be classroom groups. In particular, he should be aware (1) of some of the reasons why unit sampling may be preferable, even though somewhat less advantageous purely from the standpoint of precision, and (2) of some of the constraints on realistically possible procedural variations which the choice of the individual-sampling pattern may impose.

### Erroneous analysis of group-sampling experiments

Many studies have been invalidated by assigning different program treatments to different pre-formed classroom groups, and then proceeding to base an estimate of error on the variability *within groups* whereas the actual error based on the experimental assignment is the variation among groups. In fact, with the possible exception of some readily avoidable *faux pas* in procedure or unwitting confounding of independent variables, perhaps the *most frequent way in which the internal validity* of instructional experiments is seriously impaired lies in this error of analyzing the data, where assignment was actually by intact units, with d.f. erroneously based on N as number of individuals instead of number of units.

Even a cursory reading of the experimental literature will, further, show many instances of comparisons involving "the experimental class" and "the control class." (These terms imply a situation which--if the "class," in each case, is just one class--generally leaves no valid basis for estimating experimental error.) As Fisher notes (1960, Sec. 17, p. 35):

> It is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available. In such a case the experiment cannot be said, strictly, to be capable of proving anything. Perhaps it should not, in this case, be called an experiment at all, but be added merely to the body of experience on which, for lack of anything better, we may have to base our opinions.

To avoid this kind of situation, the instructional-media experimenter must either have access to sufficient number of groups or subjects so that he can randomly assign them to the different kinds of programs, or, if this is not feasible, must randomize the assignment of programs within individual classroom groups. The former requires him to have enough classroom groups available so that he can assign two or three groups at random to each program, using the variation among groups as the measure of error in his comparison.

When it is not possible or realistic to randomize within groups (see below) and there are not enough groups available to assign several to each treatment as class units, then we would suggest strongly that *E* simplify his design. This reflects the prejudice that it is better to get valid information on a less extensive inquiry than uninterpretable information on a more ambitious question. In any case, in the use of the unit-sampling technique the essential condition for a rigorous experimental test, as emphasized above, is that the estimate of experimental error be based on the operations that actually determine experimental error, with degrees of freedom based on the number of independent acts of randomization. Quite aside from the question of logical rigor, this means, in practial terms, that spurious attribution of effects will not result from the fairly frequent situation in which group-to-group (e.g., school-to-school) variability is grossly in excess of what would result from random sampling of equivalent sized groups.

In chapter 5 it was pointed out that the errors made in experiments can sometimes be rectified if they are recognized. The easiest locus of error to correct is of course the last--the final interpretation or drawing of conclusions. Earlier than this, however, is the often correctable error of analyzing as if *Ss* had been individually assigned, whereas what was done in the experiment (i.e., class group assignment) would require taking the class group as the unit of analysis. Where we often say correctable, we imply that sometimes it is not possible. The reasons why it unfortunately may not be possible are two: (1) The number of units is so small as to leave--even if not *no* d/f, as when there is only one class per treatment--still *too few* d/f for reasonable precision; and (2) especially when we are concerned with a past experiment rather than a current one, the data may not have been recorded, or preserved, in a form that

Note: - This page replaces the previous page 10-41

makes the proper analysis possible. The first of these possibilities will be further discussed presently in relation to the comparative precision of a unit-based and individually based statistical test. The second will be given further consideration in the later chapter and appendix dealing with data preservation and retrieval. For the present, we may note that in several cases in the literature, experimental results which remain in the unfortunate limbo of indeterminable validity could have been properly assessed after they were initially reported had the proper data been presented or preserved--namely, the data for the relevant dependent-variable measures as separately computed for each class or unit that was used as the sampling unit in the experiment. In the author's view, this should invariably be done, not only in group-sampling experiments but in all experiments using classroom groups as *administrative* units (even with individual intra-classroom assignment to treatments). (The reason for this last is elaborated elsewhere in relation to the generality of results as a function of sampling of *conditions*.)

### *Example of a unit-sampling experiment*

To illustrate the calculations appropriate from a unit-sampling experiment with several "cells" for each of two experimental treatments (see "Design II," pp. 31-36, above), we might take the numerical data presented in Table 8-1 (Chapter 8, p. 218), but [illegible]ge our statement of "what was done" in the experiment to generate these data. Suppose then, that instead of having pairs of *Ss* *individually* assigned to treatments we had dealt with intact class groups. The "pairing" of individual *Ss* would now no longer directly apply in our calculations (although "matching" of *Ss* to increase the similarity of the classes or sub-classes

(cont.)

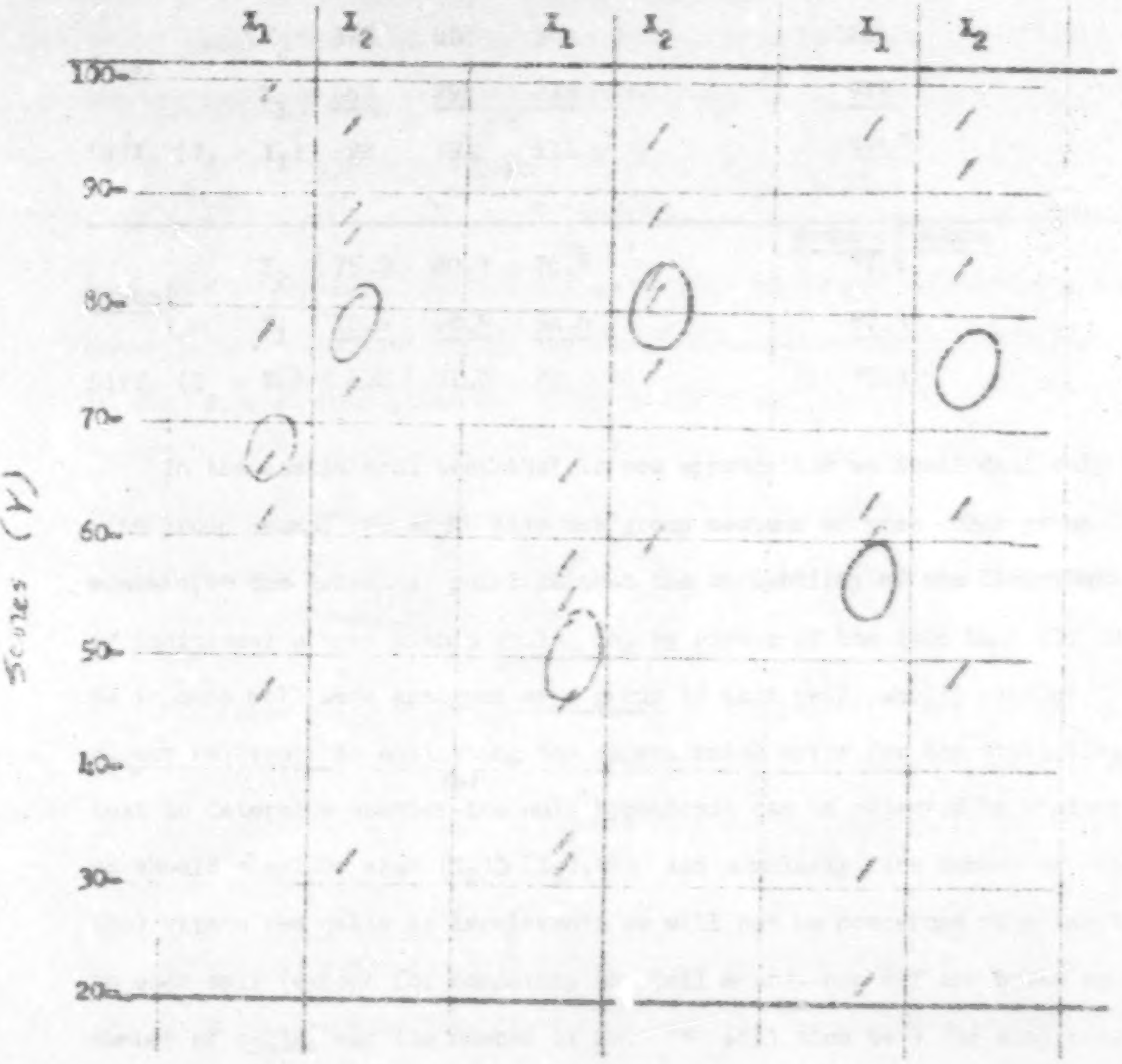
might increase the precision of our experimental comparison--see also Chapter 11, ). Instead, any pairing would have to be pairing of entire class groups as a whole and, whether we use such pairing or not, our d.f. will be determined by the number of groups (e.g., classrooms) randomly assigned to treatments, not the number of individuals.

Without reference for the moment to such pairing of groups let us merely consider that we have six groups of Ss assigned as groups to treatments in a "2 x 3" design (similar to the "2 x 4" Design II illustrated on p.10-31 above), with the scores the same as for the three $X_1$ and three $X_2$ sub-groups presented in Table 8-1. We may conveniently tally these scores as in Table 10-1.

Such tallies of scores are to be recommended, in any event--regardless of type of design--as a good way to get a look at the general nature of the data before proceeding with more abstract analysis. In this case the approximate locations of the means as estimated by a quick inspection have been indicated by the rough circles. On inspection of the tally it is evident, even before computing means, that the $X_2$ arrays will average appreciably higher in each of the three $X_1$ groups than will the sets of $X_1$ data. At the same time, however, the degree of superiority is obviously not so completely consistent, in terms of the overlap of the scores, but that some statistical analysis will be required to supplement what can be gained merely by inspection.*

The scores and means of the six "cells" of the design are, in fact, those shown below, where $Y_1$ stands for scores obtained from $X_1$ treatment-groups, and $Y_2$ stands for scores from $X_2$ treatment groups.

Table 10-1 *



* Distributions of individual S's scores (Y's) for two treatments ($X_1$ and $X_2$) assigned to six classroom groups.

| | | | | | Grand Total |
|---|---|---|---|---|---|
| Sums: | $Y_2$ | 375 | 482 | 384 | 1241 |
| | $Y_1$ | 353 | 291 | 273 | 917 |
| Diff. $(Y_2 - Y_1)$ | | 22 | 191 | 111 | 324 |
| | | | | | Means-of-means |
| Means: | $Y_2$ | 75.0 | 80.3 | 76.8 | 77.6 |
| | $Y_1$ | 70.6 | 48.5 | 54.6 | 57.9 |
| Diff. $(Y_2 - Y_1)$ | | 4.4 | 31.8 | 22.2 | 20.3 |

In the statistical test that is now appropriate we shall deal only with group means. We might also use group medians or some other group measure;** the essential point is that the variability of the distributions of individual scores within cells is, by virtue of the fact that all the Ss in each cell were assigned as a group to that cell, wholly without direct relevance to estimating the experimental error for the statistical test to determine whether (a) the null hypothesis can be rejected or whether (b) we should conclude that $(X_2) > (X_1)$.*** And similarly, the number of "cases" (Ss) within the cells is irrelevant; we will not be concerned with the "n" in each cell (except for computing the cell mean)--our d/f are based on the number of cells, not the number of Ss. "N" will thus be 3 for each treatment--that being the number of independent observations assigned to each. The d/f will, then, be either 4 d.f., if the cells were filled by randomly assigning six groups (classes, class-halves, etc.) to the two treatments, or 2 d.f., if the groups were paired with one member of each pair of groups being assigned to each treatment.**** The binomial test will accordingly not be sensitive enough; by-passing the Fisher exact test, we will compute t on the bases of both 4 d/f and 2 d/f. Obviously, our treatment means (for $X_1$ and for $X_2$) will be the same, a mean-of-means with N = 3 apiece. But $\sigma_M$ will, of course, differ for the paired and unpaired situations. For six independently assigned groups will have:

$$\sigma_{DM}^2 = \frac{2\tilde{\sigma}^2}{N} = \frac{2\left[\Sigma(Y_1 - \bar{Y}_1)^2 + \Sigma(Y_2 - \bar{Y}_2)^2\right]}{N(2N-2)},$$

where the Ys represent the basic observations (which are themselves all cell means in this case) and the $\bar{Y}$s are the treatment means (means-of-means in this case). Since N = 3, this gives us $\sigma_{DM}^2$ = 1/6 times the sum of $(70.6 - 57.3)^2 + (48.5 - 57.3)^2 + (54.6 - 57.3)^2 + (75.0 - 77.6)^2 + (80.3 - 77.6)^2 + (76.8 - 77.6)^2$, which comes out to be $\sigma_{DM}^2$ = 1/6 (276.31), or 46.05. We therefore have $t = \frac{\bar{Y}_2 - \bar{Y}_1}{\sigma_{DM}} = \frac{20.3}{\sqrt{46.05}} = \frac{20.3}{6.786} = 2.99$, with 4 d f for which P' = .02 (and P'' = .04), as shown by Fig. 8-1.

For three pairs of groups we might, in the general case, use the binomial test or Fisher's Exact test; in the present case, however, we will measure the differences by a t-test, using now only the differences in the tabulation in p. 10-43 just above.

$$t = \frac{(19.5)\sqrt{N(N-1)}}{\left[(4.4-19.5)^2+(31.8-19.5)^2+(22.2-19.5)^2\right]^{\frac{1}{2}}} = \frac{(19.5)\sqrt{6}}{\sqrt{386.59}} = \frac{47.8}{19.7} = 2.43,$$

and for 2 d.f. we find from the t-chart of Fig. 8-1 (or from a table of t in a statistics text if we don't have one of the charts handy) that P' = .07; P'' = .14.

The paired assignment thus is less sensitive than the simple random assignment without pairing. However, this will not always be the case; see Chapter 11, p. .

Reasons for choosing unit rather than individual sampling. A frequent reason for group sampling (when an experimental assignment must be in terms of units rather than individuals) is that this may be required for purposes of realism--where the kind of program treatment under study is normally used, outside of the experiment, with units rather than individuals. Or, it may be required for practical reasons, as when a limited number of schools or districts can be visited. (As pointed out by Hovland, Lumsdaine, and Sheffield [1949], a random sample of 100 members of the U.S. population would require traveling to about 100 different locations.)

Individual sampling within group situations. When administrative or other conditions require the experimental instruction to be performed in groups, random assignment of individuals to treatments is still sometimes feasible. If so, this may be preferable to block assignment of intact classroom groups from the standpoint of reducing unwanted variability due to population variables that are not randomly distributed in the intact groups. This procedure is often much more feasible in the case of two-group (two-treatment) experiments than in cases where a larger number of treatments is employed, particularly if the experimental variation is introduced only for the duration of a single lesson or class period (for example, there is often one spare room in a school that can be thus used, but not two or three). Under these circumstances, randomized assignment of the students in a given classroom to two treatment groups can often be employed, with each group then instructed (and tested also, perhaps) in a separate room by one or the other of the alternative experimental treatments. Examples of this design are found in the study of live dialog versus off-stage narration reported by May and Lumsdaine (1958, Chapter 3), in which the two thus-separated subgroups saw alternate versions of a film, and in several experiments by Gropper and Lumsdaine (1961) in which the two subgroups saw alternate versions of TV lessons presented simultaneously over two channels.*

When the experimental variations are implemented by means of printed verbal or pictorial materials, used alone or in conjunction with a constant audio-recorded explanation, within-class sampling of more than two experimental treatments, if desired, may be readily implemented within a single classroom (e.g., experiments by Kurtz and Hovland, 1953; Kimble and Wulff, 1953). Such an arrangement can be very advantageous from the standpoint of experimental economy. The same advantage obviously applies to experimentation using individually-paced auto-instructional materials. (In this case, as well as in instances where both assignment and experimental instruction are done individually, the assignment of individuals of course need not be wholly random. Some stratification or pairing may be introduced to further reduce error variance, with the experimenter then randomly assigning individuals to the treatments within pairs for broader stratification categories; see Chapter 11 )

On the other hand, when group sampling is not required, and even when considerations of precision would favor individual sampling, there is a special advantage to the "group-sampling, group-variance" design. This is that the measure of error derived from group to group variability not only takes into account the fluctuations due to individual-difference (population) variables, but also properly reflects the influence of variations in conditions of utilization for different classroom groups (e.g., time of day, instructor, behavior, procedural irregularities, etc.). A more secure basis for generalization of results is thus afforded than is obtained even from individual sampling and assignment of subjects (unless the experimental instruction ⟶

and testing is also performed individually.) By contrast, if individually sampled students are assembled for instruction and/or for testing in groups, procedural irregularities common to any particular assembled group can still represent important sources of error variance which may not be adequately randomized and therefore will not be adequately reflected in error estimates. What to do about the group in which the projector temporarily broke down during the experimental film showing is a case in point. If all, or a sizable proportion of, the subjects for an experimental condition were in this group, the experimenter is in serious trouble, and the fact that subjects were randomly assigned to that group is of little help to him. But if subjects were run individually, or if they were run in a number of groups, each of which is treated as one experimental observation, he is in much better shape; he can afford to throw out the procedurally aberrant group if the aberration is considered serious (in terms of pre-determined criteria, applied so as not to bias the comparison), or he can retain the group with the assurance that the effect of the aberration will be safely reflected in his estimate of experimental error.*

# Chapter 11

# Stratification--Replication as Related to the Precision and Generality of Data

**Stratification (including pairing or matching) and replication are aspects of structural design which are probably most often considered in relation to the precision of experimental comparisons. Here we shall consider stratification and replication primarily in this connection, but it will ~~also~~ become evident that these features of experimental structure also have an important bearing on the generality of experimental findings and on the overall efficiency of experimental stratigies (cf. Ch. 7, p. 29). The matter of efficiency applies both in the narrower statistical sense related to the power of a statistical test and also in more general respects related to the sequencing of experiments and the number of variables that may be investigated concurrently. Finally, the inclusion of more than one independent variable in an experimental analysis is closely tied to the form of experimental hypothesis being tested--e.g., to test a contingent hypothesis (see Chapter 5, pp. 41-42) will always require the consideration of at least two independent variables. However, both need not necessarily be experimental variables: cf. Chapter 7, pp. 17ff. (See further discussion of precision, sensitiveness, etc., pp. below.)**

**The implications of stratification and of the closely related notion of replication can perhaps be most conveniently introduced by recalling the elaboration of purely random assignment, involved in assigning of S's**

to treatments randomly within blocks or strata, introduced in Chapter 10 as "Design I" (pp. 31-32). In this design, available S's comprising several distinct "stratifications" or subgroups (e.g., classroom groups) were assigned to treatments (e.g., films) randomly within the subgroups or strata. This situation (which contrasts with the similar-looking "Design II," in which whole classes were assigned as intact units to treatments--see pp. 10-32ff) we will again represent in terms of a block diagram, as follows:

| subgroup/stratification | a | b | c | ... | k |
|---|---|---|---|---|---|
| Treatment $X_1$ | $n_{1a}$ | $n_{1b}$ | $n_{1c}$ | ... | n |
| Treatment $X_2$ | $n_{2a}$ | $n_{1b}$ | $n_{1c}$ | ... | n |

In this case we have turned the diagram around so that the successive classes or subgroups are strung out horizontally instead of arranged vertically for each treatment, and have indicated the possibility of an indefinite number of strata, such as k classes, rather than limiting the stratification to four categories as in the design on p. 10-31. The subscripts for the small "n" in each cell indicate that the number of cases in the various "cells" of the design need not be equal; however, for much of the discussion we shall assume equal n's in all cells.*

Sometimes the situation of randomization within cells or strata can occur simply because E is forced into this design--e.g., where he must work within the confines of successive classroom groups and because he wishes to assign S's individually to two treatments but cannot get enough S's in any single "pool" or large classroom group. Each classroom might here be said to provide one replication of the experiment; the four replications might be run off either concurrently or successively depending on circumstances (see later discussion of arguments for successive replications).

---

* Arrived at by random discarding of excess cases if necessary, etc.

On the other hand, E may elect to use this kind of experimental structure even though he is not constrained by existing institutional arrangements; for example, he might decide to obtain 20 S's per treatment giving each treatment to five S's randomly chosen from each of four classrooms--possibly 3rd, 4th, 5th and 6th grade classes (each again perhaps being thought of as a separate replication of the experiment)--instead of taking all 40 S's for the two treatments from a single large 5th grade class.

As a second example, using only a single large pool of S's, E might decide to stratify the group with four categories on the basis of I.Q., and then randomly assign S's to treatments within these stratification categories. E's reasons for these choices, in either case, might be several, as will be indicated subsequently. Regardless of his purpose, however, the formal requirements for efficient analysis of the data are the same, as will be illustrated below.

Stratification is sometimes considered separately from pairing or matching; and replication is, likewise, sometimes considered as a wholly separate topic in discussions of experimental design. In the present discussion, however, replication and stratification are in the main viewed as merely two aspects of the same essential design feature--namely, the division of experimental operations into blocks within which

S's are randomly assigned to experimental treatments. Experimentation within these blocks being carried out either concurrently or successively. (To the extent that there is a difference between the connotation of stratification (or pairing) and of replication, the former is likely to connote considerations of precision and the latter to suggest reproducibility and thus generality of results.)

Pairing

The extreme case of stratification is where the "strata" are simply pairs of individuals; for example, an efficient procedure that often is administratively workable is to obtain a class roster in advance and assign one of the numbers of each adjacent pair on the roster to each of two experimental treatments (alternative versions of a film or other instructional instrument). If the roster has been arranged in random order, assignment of every other person to one of two experimental treatments will of course result in obtaining two random halves of the group; if the roster is alphabetical, similar assignment will generally also produce substantially random halves. However, if the roster has first been arranged in order according to a stratifying variable (e.g., intelligence-test score) that is well correlated with criterion test performance, variability of the halves of classes will be materially less than random, and in this case the experimenter must of course modify his statistical analysis appropriately if he wishes to take advantage of this fact to increase the sensitivity of the comparisons. The latter can be achieved automatically when the split-class division is replicated in a large enough number of classrooms (e.g., half a dozen or so) if, despite the individual assignment of subjects, the experimenter uses the pairs of experimental-administration groups as the units of statistical analysis, as suggested above (Chapter 10) for the group-sampling case. As noted previously, this has the advantage that variance due to group-tied differences in instructional conditions (time of day, etc.) enter appropriately into the error estimate.

Statistical advantage of pairing as related to N.

We have previously seen that the appropriate statistical test when paired-observations are used is different from the test that is appropriate with unpaired observations (Cf. Chapters 10,8). The advantage of pairing is that the use of a relevant characteristic as a basis for pairing generally will introduce a positive correlation into the pairs of observations. This correlation, in general, will reduce the error estimate and hence improve the precision of the comparison provided the reduction of error variance introduced by the correlation is not offset by the loss in degrees of freedom.

By reference to the calculation of the *t*-test in the appropriate sections of Chapter 8 (for paired observations) and Chapter 10 (for unpaired observations) it will be observed that, in general, the number of degrees of freedom for paired observations is half that for an equal number of unpaired observations. When the number of observations is fairly large, the reduction in degrees of freedom will not be serious, and any reduction in variance introduced through the correlation of measures due to pairing will be benificial from the standpoint of increased precision. However, when the number of observations is very small, a moderate to marked correlation must be present before the use of pairing will improve precision. For example, in the extreme case with only two observations per group, we will have only one degree of freedom in the paired comparison and two degrees of freedom in the unpaired comparison. In this case, the correlation introduced by pairing would have to amount to around 0.6 ot 0.9 before pairing would show an advantage. On the other hand, with only slightly more observations per group, the amount of correlation needed for pairing to be the more precise method will drop rapidly. Thus, with N = 4 per group, the correlation required will range from 0.2 to 0.5 ( a higher correlation being required for pairing to be of advantage at the more stringent probability criteria, such as .005, than is required for the more limited rejection criteria, such as .05); with N of 6 or more per group, pairing will be of advantage when the correlation is more than aroudd .2 or .3, and with N of 20 or more per group, a correlation of less than .1 between the two sets of paired measures will result in a more precise comparison than an unpaired comparison with an equal total number of cases.*

Analysis of variance ~~for~~ "stratified" comparison. When strata categorized in terms of the relevant variable are used consisting of a number of individuals who are randomly assigned within strata (rather than the "strata" consisting only of single pairs of individuals), new requirements are introduced for analysis of the data. Now, in order to ~~attain the efficiency of pairing~~ ~~the comparison [illegible]~~ realize the increase in precision that stratification can provide, our comparison must employ an estimate of error that ~~ordinarily~~ accurately reflects the actual reduction in experimental error introduced by the stratification, Some form of the analysis of varience must be employed in order to accomplish this.

An illuminating characterization of the basic ~~matter of~~ nature of such analysis of variance is given by Fisher (Design, Sec. 28):

> ...The essential fact governing our analysis is that the errors due to ... heterogeneity [e.g., of subjects] will be divided, by a good experiment, into two portions. The first, which is to be made as large as possible, will be completely eliminated, by the arrangement of the experiment, from the experimental comparisons, and will be as carefully

---

* This is in marked contrast to the requirement for the size of correlation between before and after scores (cf. Chapter 9), which must generally approximate .5 before the paired before-after design exceeds the precision attainable with the after-only-comparison design (see Chapter 9, p. 26 ff.)

eliminated in the statistical laboratory from the estimate of error. As to the remainder, which cannot be treated in this way, no attempt will be made to eliminate it in the field, but on the contrary, it will be carefully randomised so as to provide a valid estimate of the errors to which the experiment is in fact liable.

For details of the rationale and computation, the reader is referred to appropriate statistical texts (see appendix A). Basically, the logic of the analysis is that several estimates of the population variance of individuals are arrived at. One estimate will be based on the variation of individuals' scores within strata. This estimate of variance will exclude the amount of ~~variation~~ variance that is ~~appropriate~~ associated with the variation among different strata, which has been controlled by the stratified arrangement. The statistical test involves a comparison of the amount of variation within strata which reflects experimental error as against the amount of variation due to experimental variables. The computation of the latter derives from the fact that the observed variation among means of two (or more) experimental groups would, on the null hypothesis, not themselves provide a basis for estimating the population variance in terms of the basic relationship $\sigma_M = \frac{\sigma}{\sqrt{N}}$.

(This discussion will be somewhat expanded in the later edited version of the chapter).

Just as in the slightly simple case of pairing, the statistical advantage of a statified design for increasing precision, as compared with purely random assigment across the board, will depend on the correlation of the stratifying variable with the dependent-variable or criterion.

### Replication

Replications of an experiment mean repetitions of it, generally with some variation in the conditions. Evidently stratifications within an experiment provide one form of replication; a different manner of replicating lies in repetition of an entire experiment, when not only the subjects differ (as they do among stratifications) but also the conditions of the experiment (including, for example, the way in which the experimental variable is implemented) may be considerably altered.

Two important points are made by Fisher concerning the use of replication. The first point (*Design*, Section 16) is that uniformity is only requisite between the objects whose response is to be contrasted (that is, objects treated differently). In this regard he further points out:

> If therefore, only two kinds of treatments are under examination, pairs of plots may be chose, one plot for each treatment; and the *precision* of the experiment will be given its highest value if the members of each pair are treated closely alike, but will gain nothing from similarity of treatment applied to different pairs, nor lose anything if the conditions in these are somewhat varied.

The second important point made by Fisher, a corollary of the first, is that:

> Paucity of homogeneous material limits the number of different treatments in an experiment, not the number of replications. It may cramp the scope and comprehensiveness of an experimental enquiry, but sets no limit to its possible precision. (Design, Sec 12).

Diverse vs. similar replications. Should replications be kept as similar as possible or deliberately made as dissimilar as possible? The answer depends on what we know (or think we know) and on what we want to find out (or try to establish). In general, there is the same kind of trade-off between generality and sensitiveness in considering heterogeneity between replications (or between successive experiments dealing with the same basic hypothesis) as there is within replications or experiments. If we want to explore or attain the maximum in generalizability, we make conditions as diverse as possible so as to see whether the hypothesis holds up over this wide range--or whether it has to be modified and translated into a contingent (rather than unqualified) relationship to account for differences in the way X affects Y under condition A vs. under condition B. On the other hand, if we have reason to doubt whether X does demonstrably affect Y in the manner hypothesized, even under previously observed condition A, then we may want to replicate under conditions as similar as possible to those previously used in order to pin down the tenability of the hypothesized relationship for at least that particular range of conditions.

An advantage of replicating an experiment under conditions as similar as possible to previous replications is that otherwise, if a discrepancy in the results of the two replications ensues with the changed conditions, we may be at a loss to interpret the findings. But an advantage of varying the conditions of replication is that the generality of a previous finding is thereby increased.

Independence of replications or repeated experiments. A particular reason for completely independent replications in which program, subject matter, learners and other conditions (including experimenter) are all changed is that possible invalidity of a previous experiment due to inadvertent sources of bias may thereby come to light. Any given experimenter working with any given subject matter is, for example, liable unknowingly to confound, with his implementation of a hypothesized independent variable, other factors not specified by the hypothesis. Merely repeating an experiment with the same experimental program will not disclose this source of spurious attribution; a wholly fresh experiment, particularly one by a new experimenter, starting from scratch with new materials, is more likely to do so, or, failing that, the greater the opportunity for the discovery of *contingent relationships*, specifying that variable X has such and such an effect under condition A but has a describably and demonstrably *different* effect under condition B.

Replication or stratification is the way we may have our cake and eat it too with respect to the competing desiderata of generalizability and precion. By using the wide range of subjects and conditions necessary to investigate generalizability, but segregating these sources of variability through stratification or replication, we achieve the condition that--while preserving the broad base necessary to demonstrate (or at least to investigate) generalizability, we are able to indentify and assess the variations due to stratifying variables, so that their influence on the dependent variable is (a) assessable rather than being buried in an undifferentiated clutter of unidentified casuses, and (b) can therefore be legitimately removed from our estimate of experimental error.

### Replication of old variables in new experiments

Aside from the increase in generalizability to be sought in replication of experiments, a special advantage is often gained in replicating, in an experiment concerned with a variable not previously studied, one or more variables whose effects have already been assessed as significant in a previous experiment. For example, one purpose of including the active-versus-passive-review variable in the Lumsdaine and Gladstone (1958) experiment on pictorial and auditory embellishments was to replicate the previous findings of Hovland, Lumsdaine and Sheffield (1949) under slightly varied conditions. This is a scientifically healthy practice that is, of course, almost routine in the physical sciences but unhappily is the exception rather than the rule in educational experimentation. There was also an additional methodological purpose in including this variation of active-passive review in a study in which the primary (new) variation was manipulation of the entertaining embellishments. This purpose was to provide a sort of yardstick against which effects of the primary variable could be assessed.

Since absolute values of the kind of scores used here have little meaning, the coexistence within an experiment of two experimental factors can give additional information that can indicate at least in rank-order terms the relative potency or importance of the two factors. This is the more informative if one of the factors is one that is known in advance to have a sizeable effect, especially if data exist (as was true in the study by Lumsdaine & Gladstone (1958) [redacted] on the basis of the previous study) as to its relative importance as compared with other variables (e.g., differences in scores associated with population factors such as ability.) In the present case we might wish to hazard the tentative conclusion that the adverse distracting effect of this type of embellishing device is of less potency--i.e., may [redacted] account for less variation--than the positive effect of overt practice provided by a good active-review procedure.

The presence of a secondary factor that predictably turns out to make a significant difference in experimental data is, perhaps, even more desirable when manipulation of the primary factor turns out to show no significant effect. With such negative results--which occur in a considerable proportion of the experimental studies in [redacted] audio-visual education, including several of the studies reported in this volume--the problem of interpretation is almost always difficult (see below.) The advantage of having the auxiliary variable that did show up as exerting a sizable influence is that this result rules out one of the possible explanations that is otherwise always an annoying possibility--namely, that the experimental circumstances were so insensitive that potent causes of real variation actually present did not have a chance to be revealed in the experimental data. This advantage is of importance in relation to the more general problems of experimental sensitivity and the logic of experiments in relation to the interpretation of negative results, discussed below.

### Replication in terms of subgroups of subjects and co-varied experimental treatments

(See Chapter 14)

An interesting aspect of methodology in the animation study by Lumsdaine, Sulzer & Kopstein (1961) arose from the fact that a factorial design was employed which in effect permitted making independent replication comparisons for different sub-groups of the audience. Thus, the animated films produce better learning than the nonanimated films regardless of whether the film was preceded by preliminary quiz which oriented men to the nature of the material to be learned; regardless of whether longer films with six instructional examples of shorter films with only three examples were used; regardless of whether or not a supplementary slide film was used following the film showing; and under all possible combinations of the three above factors. Other comparisons indicated that the animated films were superior to the nonanimated one regardless of whether the films were shown under relatively favorable temperature conditions (morning showings or under extremely humid conditions encountered in the afternoons in the experimental situation in an indoctrination training center), regardless of variations and the size of the audience and distance men sat from the screen, and regardless of whether proficiency in micrometer reading was strictly scored in terms of fully correct answers or more leniently scored with credit given for partially correct answers. Similarly, analyses were made to ascertain the differential effect as a function of the mental ability as measured by the Air Force qualifying examination. This kind of analysis is often of interest to make in view of the fact that in some cases it can be hypothesized and in some instances has been experimentally verified that effect of a particular instructional variable may differ considerably for highest function of intelligence of the learner (cf., Hovland, Lumsdaine and Sheffield, 1949, Chap.68 69.) In the case of the animation study, the effect of animation on the ability of reading micrometer settings was shown by a difference of 72% for mean correct answers for the animated films versus 66% for the nonanimated films, in the case of the more intelligent men (airmen qualifying exam, 60 and above), whereas for the less intelligent men the mean scores were 41.1% correct for the animated films and 28.3% for the non-animated films. Analysis of the data for learning of nomenclature showed that the beneficial effect of the animation devices used was also consistent for groups of trainees differing in intelligence and educational level.

## Generalizability versus Precision as a Trade-Off in Experimental Design

The generalizability of an experiment is, in general, favorably affected by heterogeneity of learners, tasks and conditions: The more various these are, the wider a range of applicability the experimental results may be shown to have. On the other hand, the narrower the range of variation with respect to conditions and ability of learners, the more precise and sensitive the experiment may be. The narrower the range of variability in learner ability, the more influence in the final outcome a potent treatment variable will make; wide variation in learner ability tends to increase the size of the error term and to prevent effects of the independent variable from showing up as "significant"--i.e., demonstrably non-chance--differences. Here again, Fisher's statement (*Design*, Sec. 10) is cogent:

> The method of pairing, which is much used in modern biological work, illustrates well the way in which an appropriate experimental design is able to reconcile two desiderata, which sometimes appear to be in conflict. On the one hand we require the utmost uniformity in the biological material which is the subject of experiment, in order to increase the sensitiveness of each individual observation; and, on the other, we require to multiply the observations so as to demonstrate as far as possible the reliability and consistency of the results.

The question of external validity or generalizability is essentially a question that asks to what universe the sample of data obtained in the experiment applies. This involves, unfortunately, some far from completely solved problems. Two contrasting approaches may be taken. The first is (given that internal validity is granted) to put the burden of proof on the skeptic--that is, to assume that the results are generalizable to any subjects and to any group of learning tasks whatever, unless proved otherwise or unless there are good reasons for believing otherwise. The opposite approach is to assume that the results are applicable only to the population from which one has explicitely sampled. (For example, one might have defined a group (a "population" or "universe", in the statistical sense) of all possible quadratic equations, or of all relevant numerical computations, or of all English words, and would then limit his claim to this population by working with an explicitly-drawn sample thereof.)

Obviously, there are many instances in which one is unwilling to do the former and either unwilling or unable to do the latter--unable because the population can't be defined rigorously, and unwilling because it doesn't seem reasonable to admit that the applicability of a result is limited solely to, let us say, the learning, by Miss Hortense Jones' 5th grade class at Bagley School on the 22nd of June, 1961, of the list of English prepositions found on a given page of the 6th edition of Hiram Webster's textbook. The relation of generality to theoretical formulation of problems for inquiry is discussed in Chapter 12.

Precision in experiments

The precision we will desire of an experiment depends on our estimate of the magnitude of the differences we wish to detect. If we are dealing with factors that we think are very potent, a relatively small or unprecise experiment, provided it is valid, will suffice. On the other hand, our concern may be with a factor which, while of considerable potential consequence in the long run, is not expected to produce huge effects. If our experimental purpose is to test whether such a factor does in fact have any effect we will need a relatively precise experiment, or more broadly, a relatively sensitive experiment.

It is pointed out elsewhere that there are various ways of increasing the sensitiveness of an experiment. These include increasing its size so as to get more stable values in which the contamination from chance fluctuation is ironed out to a greater extent. Also available is a variety of techniques for increasing the sensitivity of the measures; for example, using a longer test or an individual testing technique which may yield more stable or sensitive measures than a group test. In addition, we may help to reduce variability due to individual differences by various techniques of matching or pairing--that is, choosing a structural design for the experiment which is conducive to precision.

Given a particular set of experimental results or a particular design, we may also have some choice as to the kind of analysis to be performed. Both in estimating the size of an experiment and choosing among statistical analysis with differing degrees of power, a guiding consideration can be the following: If a given size of experiment or precision of analysis will suffice for our purpose--that is, will permit us to reject the null hypothesis at a desired level of confidence, thereby demonstrating that the factor in which we are interested can be concluded to make a difference, we need not have recourse to a more precise experiment or a more precise analysis. If we are dealing with a relatively potent effect, a relatively small experiment will also suffice and it can be well argued that in this case the conduct of a very large experiment is a waste of resources. If we are satisfied to conclude confidently that a factor makes a difference when we reject the null hypothesis at a confidence level of .01, little is gained by increasing the precision of the experiment so as to improve the confidence level to a value of .0001. The same is true with a statistical analysis. If a relatively inefficient analysis (statistically speaking) will suffice to reject the null hypothesis with experimental data which in fact we have at hand, there may be little point in conducting a more refined and laborious analysis.

It is true that, ideally, had we been able to estimate accurately the amount of effect and amount of variability in the data in advance, we might have designed a smaller experiment. But we cannot always make this prediction accurately--indeed, our inability to predict exactly what the effect of the experimental variable will be is indeed often our primary reason for doing the experiment in the first place. For this reason, we may often wind up with experimental data in which we need not go through the statistical labor of extracting the full information from the data. For example, suppose we have obtained a set of observations in comparing two treatments, $X_1$ and $X_2$, which have been administered to eight pairs of individuals (or pairs of classroom groups), and have obtained all eight differences favoring $X_1$. All eight of the directions, regardless of amount, have positive signs--that is, all eight differences are consistent in the direction predicted by the hypothesis--reference to the binomial probability chart or table (see Chapter 4) tells us at once that the null hypothesis can be rejected at the .01 level of confidence without even bothering to calculate the statistics required for the more powerful t test; and it can be argued that in this case, the labor of doing the latter might as well be saved. If, on the other hand, given the same data in which one of the signs is negative, and if we want to have the assurance of the stringent confidence level of .01, then the binomial test will not suffice, and we must resort to a more powerful statistical test.

The test we will choose might be determined both by the labor involved and by our concern with whether or not assumptions underlying a more powerful test, such as the t test, can be regarded as met. A somewhat more powerful test--that of Wilcoxon for signed ranks--can be applied with a minimum of assumptions and labor. If such a test leads us to reject the null hypothesis at the .01 level, the fact that the t test might give us a still lower probability value may be of little consequence.

The amount of statistical labor involved in conducting tests may be relatively slight even for more powerful tests such as the t test or analysis of variance in some instances. But if we have a large number of tests to conduct we do have to balance off the cost of the clerical labor necessary to conduct and check them as against other possible ways of increasing the precision of the comparison. In some instances, indeed, it is more economical to run a larger experiment where the data can be cheaply gathered than to use a more refined statistical analysis. (Cf. The interesting commentary on this point is made by Mosteller, 1948, and by Mosteller and Bush, 1954.)

In any case, the experimenter is generally curious to find out some indication of the outcome of his experiment as quickly as possible and it is often of interest, as soon as the relevant data with respect to direction of differences can be obtained, to apply a quick and inefficient test which may, as we have seen, in fact suffice to satisfy the primary purpose in which he is interested. It can also be argued that the conduct of a relatively simple though inefficient test is a good check on the possibility of arithmetical errors which are more likely to occur in the more elaborate test. The habit of using a simple test like the sign test is also regarded by the authors as a healthy practice because of the fact that such a test involves a minimum of behind the scenes statistical manipulation, and keeps the process of the experimenter's drawing conclusions as closely tied as possible to direct steps of reasoning which he can fully apprehend. Such a habit is also useful in reading of research reports by others where inspection of the data from the standpoint of a simple test like the sign test will often provide a quick and rough, but quite adequate, check on the significance of the differences which the author of the report has arrived at by a more laborious though somewhat precise method.

Unfortunately, the essential data needed to check the conclusions ground out by the author's statistical mill often are not reported. It is the belief of the authors that the essential data necessary to substantiate the conclusions should almost invariably be included in a research report, either in the text of the report or in a supplementary table. Where to draw the line in this respect is often a difficult decision, however, and one would not expect that the very "rawest" of raw data would be completely presented. Rather, some intermediate form of summarization, such as the relevant scores for each subject in an experiment involving individual assignment or the relevant class means in a group assignment experiment would be presented.

Other factors in the sensitiveness of experiments

Aside from lack of precision due to faulty design or analysis, an experiment may lack sensitiviness because one or more of the following occurred:

(1) The implementation of the experimental variable was lacking in sufficient efficacy to produce a sizable impact or effect. (In this case, effects *if* demonstrated as uncontaminated and non -chance have to be accepted; but *potential* effects actually implied by the hypothesis may may fail to be revealed because the experimental treatment was weak [lacking in efficacy] in relation to the potential potency of the hypothesized casual factor).

(2) The measuring instrument lacked sufficient sensitivity to reveal effects that actually took place. (e.g., there may have been a difference between how much was learned from two programs, but the learning wasn't fully measured and hence didn't show up.)

(3) Even though a treatment actually succeeded in producing a difference in the dependent variable, the null hypothesis cannot be rejected because it is impossible to demonstrate that the obtained difference was not a chance result. This might be reflected in a fairly sizable mean difference between two treatments but with, also, large uncontrolled variation in learning or achievement, so that the estimate of experimental error is large in relation to the difference between achievement for the two treatments. It might also result from the use of an inefficient statistical test--one of lesser power than is permitted by the available data.

The difference in status of these three cases may perhaps be clarified by relating them to what would happen if the experiment were to be replicated or repeated, assuming that in fact the hypothesized factor is one of some potency. In this case, statistically "significant" differences--i.e., differences demonstrable as non-chance at an acceptable level of confidence--could in case (2) or (3) be revealed with the same treatment, merely by increasing the precision of the experiment. In doing so the experimenter might (1) find it more efficient to increase the sensitivity of his measuring instrument (e.g., use a test with more items, or with more discriminating items), or (2) might find it more efficient to use to use the same measure but increase the structural precision of the comparisons. The latter might be done either by increasing the size of the experiment (testing more students or more classes) or by using a more precise design (e.g., by matching or pairing students or classes on the basis of ability). Also, he might (3) tighten up procedural controls so as to reduce the chance variation due to extraneous factors.

Efficacy of independent variables and pre-test tryout of programs.

Lack of efficacy (or of a "green thumb") in implementing an experimental variable in an actual program results in obscuring the effects of actually potent sources of variation in a statistically non-significant differences. This situation often stems from the premature conduct of formal experiments prior to adequate preliminary experimentation.

The efficacious operation of most experimentally manipulated factors require an extensive de-bugging process as well as skill in constructing experimental materials that potently exemplify the operation of the specified experimental variable. Perhaps the commonest example is found in the conduct or premature demonstration experiments to show the value of a new general "method" before the materials presumed to exemplify the method have been subjected to sufficient development and refinement to insure that a specifiable method is adequately exemplified (see below). However, the same fault is often apparent where a more specific hypothesis about a particular instructional factor is presumably being tested.

Another question concerns the duration or length or magnitude of the instruction that is experimented with. Many experiments use very short programs in the range of half an hour to an hour. By contrast, in such experiments as that of Porter (19__) or Klaus and Lumsdaine, the forms of experimental instruction compared extended over a period of several months or several weeks.

The difficulty and the cost of conducting an experiment may or may not go up more than proportionally as the duration of the experiment increases. But methodological difficulties of conducting a cleam experiment tend to go up very drastically as its duration increases. This is not just a continuous simple, straight increase, but goes by jumps. For example, if an experiment is one that can be conducted within the course of an ordinary class period, we have a captive audience for that period of time. We then can get our initial measure, present the experimentally varied forms of instruction, and get a terminal measure, all while we still have our objects all in the room. The possibilities of conducting a well-controlled experiment are vastly greater if we have to finish the experiment even the next day. We don't know what has happened in the meantime to possibly contaminate the results.

Suppose we have an experiment comparing two forms of teaching machine programs. If we do this all within a class period, we know pretty much what has happened. But if we take the measurement the next day, or over several days later, it is possible that some of these students who have one form of program may for some reason do some auxilliary reading on the side, and there may be a differential.

## Quasi-Experiments and Matching

In a true experiment the extimate of experimental error is generated directly by the number of independent acts of randomization. Sometimes the number of experimental units available is considered too limited to achieve this condition, and other means are resorted to. The logic here is an "as if" logic: We try to achieve a condition wherein the difference between two groups assigned to two treatments can be treated "as if" the members of the groups had been assigned at random. We can never really achieve this "as if" condition, and hence our estimate of error will always be wrong. Since we ordinarily want to insure that we do not erroneously reject the null hypothesis, the usual tendency is to try to match the two groups so that the difference in their composition will be at least as small as would have been obtained by random sampling, and then use a statistical test which assumes random sampling. In doing so, we may say we err "on the conservative side"--that is, we hope that differences which the statistical test indicate as dependable or "significant" will be "true" differences, even though we may fail to demonstrate some truedifferences that actually exist.

In such a quasi-experiment, comparability is generally achieved by some form of matching or equating--either by physical equating through selection or selective discarding of cases, or through statistical means such as covariance analysis (where a further "as if" is introduced--we treat the groups as if they were comparable or matched, where in fact they are not).

It may be noted that several methods may be used within a given quasi-experiment to try to achieve comparability and to try to achieve a condition where a statistical test, demanding assumptions which have not in fact been met procedurally, will yield results which there is reason to think will, though inevitably "off" to some extent, and inevitably in violation of the requirements of rigorous logic, be at least "in the right ball park."

A quasi-experimental analysis is indeed sometimes performed where the conditions of a true experiment were present and could have been exploited. The conspicuous example is on the case of a group experiment where a sufficient number of class-groups were assigned randomly to treatments to provide a reasonably sensitive and still quite rigorous statistical test but instead an invalid test based on within-group vaisbility was used.

One way of attempting to achieve comparability (eliminate bias) by use of equating or by co-variance analysis employing background or prediction variables to match the groups so as to assure that they are "equivalent" except for the effect of the experimental variable(s). When matching is to be used, matching the groups by selective discarding of cases may be less efficient than use of co-variance in the sense of purely statistical efficiency, but the statistically less efficient method of equating may be preferable nevertheless, especially where large numbers of causes are available and when, as is not infrequently the case, there are a number of prediction variables in terms of which it appears relevant to match the groups. It is usually preferable to match groups by distribution categories rather than solely in terms of means. It should be noted here, however, that groups may be grossly dissimilar with respect to certain background variables without affecting the outcome of these variables are uncorrelated with--and hence irrelevant to--the criterion variable (see Hovland, Lumsdaine, and Sheffield, p.324-25). Use of selective-discard matching (to even up the distribution with respect to several stratifying variables) rather than employing co-variance may save labor (at least unless the volume of analysis warrants programming for digital computer calculation) when the number of dependent criterion variables is large, as in the case of detailed evaluation studies (e.g., Hovland, Lumsdaine, and Sheffield, 1949, Chapters 2 and 3; May and Lumsdaine, 1958, Appendices A and B), since one process of equating suffices for any number of dependent variables (points of information to be conveyed by a particular instructional film). For efficiency, it is necessary, of course, that the matching variables be of general relevance to this whole class of criteria.

Matching (or use of co-variance) to equate pre-formed groups may, if carefully carried out, result in demonstrably close "comparability" of the groups to be compared (e.g., in terms of very close similarity on a relevant pre-test), but sole dependence on its use to compensate for lack of individual sampling still leaves the experimenter without a rigorous basis for assessing experimental error that derives clearly from the operations performed in the assignment of experimental treatments: The investigator who would use matching as a sole basis for "comparability" runs the hazard that he must assume that the matching results in group composition that does not differ materially from what would have been obtained had random assignment to experimental treatments in fact been employed. Such "let's not and say we did" procedures can never be as satisfying from the standpoint of experimental rigor as can procedures in which the way in which experimental treatments are assigned directly generates the method of computing the measure of experimental error and the test of significance that is used to determine whether the differences in criterion measures associated with differences in experimental treatment represent dependable, non-chance effects of the treatments. Accordingly, the former procedures should be resorted to only when compelling administrative factors preclude the setting up of an experiment in which the unit entry that contributes one degree of freedom to the error estimate (be it a person, a classroom, a school, or a city) in the statistical analysis has a one-to-one relationship to the unit of random assignment by which subjects are allocated to experimental treatments.

# Chapter 12

# Strategy and Tactics in Instructional-Media Experimentation

## Research strategy: choosing problems for investigation

As noted at the outset of the volume, our realm of discourse lies within the field of experimental studies of the effects of media. The book has thus been addressed primarily to experiments capable of testing hypotheses about factors having some degree of potential generality--as distinguished from the mere testing of products (to determine, for example, the effects of a particular film, a specific self-instructional program, or a given tape-recorded lesson). As we have also pointed out in Chapter 1 (p. 31 ), however, many of the basic points of method which we have discussed in describing how to do experiments to test hypotheses apply also in varying degrees to specific "technologically" oriented studies designed to assess particular products rather than contribute to principles or to the building of a science of instruction. Specific aspects that in the main apply only to the assessment of effects of specific products without regard to the derivation of general principles fall outside the scope of the present work.*

## "What to do" versus "How it's done"

So far, in parts II and III of this book, we have talked mainly about the tactical questions of "how to do" an experiment to test a hypothesis about the effect of instructional media. In doing so, we have made only occasional reference to the complementary strategic question of what experiments to do, what instructional-media hypothesis to test. Another way of saying this is that we have talked about the formal structure and procedural aspects of experiments, and in doing so have described the manipulation of independent variables without much discussion of what it is that is manipulated. In other words, we have dealt with ways to test hypotheses without much examination of the content or reasonableness of the hypothesis to be tested.

However, it is evident that we not only need to view with a critical eye the methods by which an experiment is conducted but must also try to become sophisticated about what it was that was being experimented with. What instructional-media research to do is surely as important as how to do it. It is hard to attach greater importance to either: One can not say either that the selection of the hypothesis--that is, basically, the question to be answered--is more important than obtaining a sound answer to it, or vice versa. Ask a foolish question (i.e., start with a poor hypothesis to test) and you will get a silly answer--no matter how impeccable your experiment in terms of methodology. The goodness of the question asked is thus a necessary condition for fruitful research, though not a sufficient one. But obviously, the same is true about the goodness of the methodology employed. It is a necessary condition for a successful experiment to have sound methodology--with which we have mainly been dealing in the preceding chapters. But to apply painstaking and careful methodology to a question that is hardly worth answering anyhow is clearly good effort gone to waste.

On the other hand, an erroneous answer (in terms of experimentation, a conclusion which is unsound either because of poor methodology or inaccurate logic in drawing the interpretations from data) might in some ways be considered even more deplorable since here, if the fallacious nature of the conclusion is not detected, mistruths may be perpetuated under the assumption that they are in fact true. Thus, rather than errors of omission in the building of an instructional science, we have errors of commission--so that the literature gets cluttered up with undependable generalizations parading as established knowledge. Thus, no matter how interesting, important and well-formulated an experimental hypothesis is (that is, no matter how good the question you are asking), the question of how you answer it (how you do the experiment) is also crucial if we are not to be grossly misled. These facts may seem self-evident but are made explicit here as we turn, in this and the next two chapters, from questions essentially dealing with the "how" of experimentation to a consideration of factors bearing on "what" to experiment on.*

## Approaches to the Question of "What Research to Do?"

If both the "how" and "what" of research are important, it is certainly true that both also represent very difficult matters. We have already seen this with respect to some questions of "how" in Chapters 3-11 which, we trust, have helped to document the warnings of Lavoisier and others concerning how great are the difficulties with which experimentation is fraught and how important is the basic understanding of each step of the experimental research process, if experiments are not to be unsuccessful. But the bases for possible lack of success in experiments with which we have dealt have, thus far, had little relation to the kind of question asked as distinguished from the methods used to answer it. The authors believe that, despite the "how", difficulty (of achieving an experiment with thoroughly sound methodology), the "what" question is much harder to answer than the "how" question. Indeed, as indicated at the outset of the book, we have little assurance that we can help the reader much in the selection of research problems. On the other hand, as earlier indicated, there do appear to be some suggested guidelines which may help.

What instructional-media research should you do? How do you go about finding a sensible hypothesis--an important question the answer to which will make a positive even if small contribution to the building of an instructional science?

The answer to the question of what research to do will, of course, depend partly on your interests. But also it depends on some consideration about what is a sensible question to ask--that is, a question that is answerable and one to which the answer will make a difference. Let us first consider the bearing of the investigator's interests on what research should be undertaken.

## Interests of the investigator as a basis for selecting research problems

The comment that the answer to what research on instructional media you should do depends partly on your interests is based in part on the belief that it is clearly impossible, at any stage in the history of inquiry, to identify the crucially most important question(or, at least, to obtain agreement on this). That the interests of the individual investigator will to a considerable extent determine what problems he works on is thus desirable, in the absence of omniscience, even though it necessarily involves some wasteful duplication and overlapping, and also some lack of integration in bringing together common elements of diverse problems in a single programmatic effort. In effect, the result is like the strategy of multiple salients in an advancing battle front, in which effort in direction A may subsequently be withdrawn or redirected depending on something that has happened at another sector of the front such as a breakthrough across an important river. But if we stop and wait to know what all the progress on other fronts might be before making attempted progress on each of a number of other fronts, the total process, though conceivably more efficient, would obviously take much more time. Thus, more rapid progress may be made by a variety of concurrent attacks, even though some of these will inevitably turn out to be less successful than others.

The fact that in hindsight we could have said it would have been better for Experimenter X to work on Hypothesis B than on Hypothesis F (which he did work on) does not mean than any master plan for research is possible in advance. On the other hand, it is certainly possible to direct interests of investigators in making salients across a broad research front in directions which, on the basis of past experience, seem relatively likely to succeed or less likely than others to fail. (Indeed, this is the intent of most of the book from the standpoint of methodology, and of the present section from the standpoint of content of research.) The position that we will take is that certain directions of effort have inherent limitations already identified sometimes on the basis of extensive and costly past experience which make them almost certain to fail, whereas others seem, tentatively at the present time, to be more likely to contribute to real progress.

Feasibility and expediency. There is always a factor of expediency in any experimentation. The extreme form might be where one has a film and asks, "Now, what experiment could I do with this film?" But even in the ordinary case where we have first identified the problem for investigation, we must next look for a feasible way to conduct the experiment, and what we do is generally some kind of a compromise between what we would ideally do, and what is practical. What can one afford financially? What can he get done by the time he is supposed to turn his thesis in? What can one persuade his colleagues to accept? What effects are we able to assess?* The choice of the particular subject matter of a particular experiment is thus based on many considerations, some of which are lowly, down to earth, mundane considerations, and only some of which are based on logical criteria.

## Levels of Decision in Educational Research

In dealing with the choice of research problems, it may be helpful to make a distinction between levels of decisions. The distinctions we would make in relation to the "what" and "how" questions in research are primarily those between the strategy of experiments and their tactics. However, two more or less supra-strategic levels of decision can also be identified --namely, educational philosophy and research policy. The reason for making these distinctions is that, without them, confusions are likely to arise, and issues are likely to be settled on one ground or one basis, de facto, when it is assumed that they are being settled on some other basis. (This is also true, par excellence, of the distinction between the validity and the precision or sensitiveness of experiments.)

### Educational philosophy and research policy

Educational philosophy. At the philosophic level is the prior decision to concern oneself with education rather than with, say, medicine or politics. Decisions at the level of what we are calling educational philosophy then are posed by such value questions as the ultimate aims of education.* Such decisions have implications for research strategy in that they may govern the choice of what to experiment on. This is particularly true to the extent that the identification of a research problem or hypothesis for investigation is formulated first in terms of a choice of dependent variables--that is to say, the kinds of instructional objectives which it is sought to achieve.* Obviously, a philosophic or value decision may place more importance on studying how best to realize certain kinds of instructional outcomes rather than how best to realize others.

Research policy At the level of what we are calling research policy comes such questions as: Given that we have already selected the improvement of instruction and, (perhaps, certain kinds of instructional objectives), as our focus of attention, a question of policy now becomes whether to seek increased knowledge, improved products, or wider utilization--that is, whether to do research, engage in program development, or bend our efforts toward dissemination of research and development of products (knowledge, techniques, or specific programs) already developed or discovered.** Do we want, in short, to do more research aimed at gaining further knowledge, improved techniques, or better products--or to work at insuring better utilization of the knowledge, techniques, and products already in inventory on the basis of past research and development.***

Research strategy and tactics

Assume that the policy question has been settled--that is, that we have determined the kind of general or specific educational goals to which we are going to seek to contribute and have decided that the means by which we are going to seek them is through some form of research to increase our knowledge about the effects of instruction. The question now is what research to do?

A question of grand strategy here is whether to use experimentation or to rely on non-experimental methods. Also at this level of decision, is whether to do scientific experiments that involve primarily the acceptance or rejection of hypotheses of some generality, or whether to do experimental research with "evaluative," technologically oriented, or product-testing aims. A related question is whether we shall be involved in estimation of parameters or only with hypothesis testing (see discussion of these characteristics elsewhere). By the terms of reference adopted in introducing the book we have already made our choices with respect to policy level and grand strategy; that is to say, we are committed primarily in this work to dealing with methods for experimental inquiry and, particularly, with scientific experiments for hypothesis testing, with only incidental implications, as a by-product, for the methodology of evaluative experiments, and with only limited concern for questions of estimation.*

*Major strategic decisions in experiments with a scientific purpose.* Given the decision of grand strategy to do scientific, hypothesis-testing experiments primarily aimed at increasing knowledge that will contribute to a science of instruction as a basis for future decisions, one major strategic decision involves the choice of dependent variables--that is, instructional objectives of a broad sort--or, at least, the question of whether to work on general principles appropriate generally to all learning, or those specific to particular kinds of instructional objectives such as knowledge or skills or specific to indoctrinational objectives such as persuasion.

A further major strategic question is now the questions of what *general kinds of* independent variables to investigate and, related to this, what *kinds of* hypotheses to test; also, what degree of confidence to seek in the results.

In addition to these choices, we have, prior to the further strategic decision of choosing what specific independent variables or hypotheses are to be tested, to choose between theoretical and atheoretical formulation of the hypothesis, and to decide what complexity of hypothesis to concern ourselves with. The major decision here is whether unqualified or contingent generalizations are sought (cf. Chapter 5, p. ). If the latter, a multi-factor design of some kind is implied, and thus the determination of whether to use the very simplest experimental structure or one which is at least somewhat more complex is determined at this level by strategic decision in terms of the form of the hypothesis to be investigated. (Later decisions of a tactical nature also affect the complexity of the experiment in terms of structural considerations affecting experimental precision , logistics, and efficiency--in terms of information gained for a given expenditure of resources or cost.)

Minor strategy or major tactics. Here we have the question of the kind of general structure of experiment to be employed, and the general methods to be used for insuring the internal and external validity of the experiment. Beyond this we will consider the specific ways in which the experiment is conducted, particularly with respect to how to achieve a given degree of precision or sensitiveness, as primarily questions of tactics.

Also in the field of tactics would appear to be the choice of procedural patterns of experimentation (e.g., before-after versus post-test-only or after-only forms of experimenting), the choice of whether to control by stratification, sampling of subjects versus of subject-matter, and the like. It is obvious, however, that ultimately decisions on tactical matters may be governed by, and their outcome may contribute to, further strategic decisions--for example, the efficiency of an experiment ultimately determines how fast we can progress with a given amount of resources, or what level of resources we try to obtain in order to conduct our research.

### Strategy, tactics and "design"

As noted in Chapter 5, Tukey (1960 ) has made the point that the term design, as applied merely to the formal structure of an experiment, has too narrow a meaning. We share this view and follow his conception is using "design" in the general sense of total research design, from selection and formulation of problem through the determination of structure, the planning and implementation of procedural controls, and the methods of tabulating and analyzing the data (insofar as the analysis is separable from the structure of the experiment). Design may thus be said to cover both strategic and tactical aspects of experimentation.

When we distinguish between the strategy of an experiment and the tactics employed, we use these terms in much the same sense as in military operations (we could even think of strategy and tactics as being employed, if we wish, in a war against ignorance). The strategy of experimentation involves, we might say, what experiment to do whereas tactics involves the question of how to carry out the experiment. Strategy is thus the purpose of the experiment and tactics the method. However, the differentiation is not this clear; and there is a considerable area of minor strategy or major tactics which concern the validity of the experiment.

As a resolution of the overly particularized meaning given to design in many treatments of experimentation, we can, as indicated above, best use the term "design" in a generic sense. The term "structural design" or structure of the experiment can be used to indicate the narrower sense which Tukey calls pattern (we might call this "structural pattern"--or, following Fisher, simply use the term "structure"). Structural design can be distinguished on the one hand, from procedural design [or procedural pattern or procedural plan (see Chapter 9)] and, on the other hand, from conceptual design, which refers to the strategic considerations involved in the choice of hypotheses and the general way in which hypothesized factors are translated into experimental operations. Structural and procedural aspects of design primarily determine the validity (as well as the precision) of the experiment--and in turn determine the kind of statistical analysis or test of significance or confidence which will be appropriate.

## Strategic Choices as Related To Type Of Independent Variable

Hovland, Lumsdaine, and Sheffield (1949) have made a useful distinction between the "evaluation" studies which asses the effect of a single film (or other single program) and the evaluation of a class of programs, which may be taken as representative of a particular medium.

The experimental question posed here may be simply: How much did a particular instrument (e.g., film) teach? This requires a way of determining the "before" and "after" status--and hence the change from before to after--of those who were given the instruction. Alternatively, if the evaluation is one in which the effects of two instruments are compared, a comparison of the "after" status resulting from the use of each will suffice for some purposes (provided initial comparability of the two groups of students is assured), though a more complete picture is, of course, provided if a "before" baseline is also provided. (See Chapter 9)

An important variant of this kind of comparison is one in which the attainments effected by a particular instrument are compared with those obtained by some alternative form of instruction characterized as "conventional" or "currently used" instruction--where the latter is not mediated by a specifiable single alternative instrument. The severe restrictions on interpretation of such a comparison arise from the lack of specificity of the instruction with which the instrument is compared; the same kinds of restrictions apply to a considerable degree in the comparison of two classes of media, and are discussed subsequently in that connection.

The restrictions on interpretation which inhere in evaluative studies can be identified by considering the purposes of such studies. These purposes include, first, the basis for some over-all assessment of the instrument--to determine, for example, whether anything has been learned through its use, or to determine, in comparison with some implicit or explicit standard of reference, or basis for expectations, whether enough is learned through its use to warrant its adoption for some future instructional purpose. Such a decision can be made on the basis of an over-all assessment of effects--for example, total amount of gain on a relevant test. Where the evaluation made in over-all terms is a comparative one involving two alternative instruments or presentations, the purpose may be to determine which of these two specific instruments (e.g., two specific films) to adopt. (A legitimate purpose is not the determination of whether similar instruments of the same general class should be accepted for further use in the future: see below.)

## Diagnostic evaluation

Aside from over-all assessment in terms of some total score, a very important class of evaluative experiments are those conducted for "diagnostic" purposes, in which specific effects of the instrument on a number of specific points are separately measured. If the instrument is an orientation film, for example, these might include points of factual information and a variety of specific opinions which, it is hoped, will be influenced by the film. Here one would pay attention not only to the total score but to the amount of information learned on each specific test question or point. To achieve adequate sensitivity or stability of results, this diagnostic purpose may, of course, require a larger sample than would be required to detect differences of the same magnitude in an over-all score.

Data on the instructional effects achieved (and on those of the desired outcomes which fail to be achieved) by the use of a particular instrument are useful ~~primarily~~ primarily in deciding whether to use or further develop a particular instrument. If the instrument is one which had been completed (say, a completed and released instructional film), then the use of information is limited to indicating whether it is worth while to use the instrument and, if so, what specific additional instruction may be needed to achieve the defined goals. Such results can, for example, show the teacher or curriculum supervisor which points of subject matter to devote special attention to in subsequent classroom instruction, as distinguished from those which can be achieved simply by showing the film.

One of the most important uses of such diagnostic evaluative experiments is realized when they are carried out on a preliminary version of an instrument, or on an instrument which is to be revised. The results can then be used in modifying or redesigning the instrument. As pointed out by Hovland, Lumsdaine, and Sheffield (1949), such studies should be carried out early in the stage of production and repeated after each major stage of revision so that, by successive correction and re-evaluation, the desired instructional goals can be achieved in a satisfactory manner. In the case of group media of instruction such as films, such evaluative testing and successive revision has long been advocated (e.g., Lumsdaine, 1947; Carpenter, 1948) but has seldom been carried into practice. One exception is the recent series of studies of television programs by Gropper and Lumsdaine (1961). This approach is also being pursued vigorously in the development of materials for use in teaching machines or as "programed text" material.

An Air Force study reported by Zuckerman (1954) explored the possibility of using preliminary filmstrips, based on "story board" sketches, as media for predicting the effectiveness of films prior to their completion. In this study, a comparison was made between a preliminary filmstrip with accompanying narration and the finished sound motion picture, "Flight Capabilities of the F-86A." After viewing the filmstrip or the film, the subjects (Air Force pilot students) were tested on specific points of information presented in the films. The data were analyzed in terms of the performance of each group on each test item. The results showed a high degree of correspondence between percentages of answers given correctly by the film group and by the preliminary filmstrip group on each point of information--e.g., those test items best learned from the film tended also to be the ones best learned from the filmstrip. This correspondence between film and filmstrip was also present for the more difficult items. Based on all test items, the correlation between percentages of correct answers for the two groups exceeded .80.

Although Zuckerman's study established the feasibility of predicting the degree to which different points would be learned from a finished presentation, it did not go on to demonstrate how such data from a preliminary form of presentation could be used to improve a revised version. This was done in a later study by Gropper and Lumsdaine (1961).

Two lessons, one on "heat" and one on "chemistry," were prepared by two different television instructors. These lessons were scheduled as a part of a regular junior high school science series. Before being shown on the air to the entire population of classes that ordinarily watch the lessons, a preview showing of the "heat" and "chemistry" lessons was transmitted (on a UHF channel during nonpeak hours) to a few sample classes. Following the program, students were given a quiz to determine which points in the lesson had been communicated effectively and which had not. As anticipated, item-by-item analysis of the test results indicated that, while certain points in the lesson were fairly well understood, certain other points were much less well learned. The television instructors then revised their presentations to attempt to remedy the specific weaknesses of the lesson as revealed by empirical data.

To test the improvement in effectiveness resulting from this try-out and revision procedure, the original and revised versions were telecast simultaneously to eight matched classes over WQED-WQEX, a unique two-channel educational facility in Western Pennsylvania. Following the simultaneous telecast, students were tested on their comprehension of principles as well as their knowledge of facts presented in the lesson. Students who watched the revised lesson tended, in general, to acquire a better understanding of the concepts presented. This was true for both the "heat" and the "chemistry" lessons. Students who watched the revised version of the heat lesson averaged 9.5 points higher on a 90-point test than those who watched the original. For the chemistry lesson, the students who watched the revised version averaged 12.9 points higher on a similar test than those who watched the original. Both differences were statistically significant.

The results of this study support the notion that student responses on achievement tests administered following a preliminary showing of a lesson may be used to provide useful feedback to the television instructor. The information about students' responses helps the instructor to adjust his presentation to the level of his audience, thus providing a substitute for the "live" responses he might get in a small classroom. Student test responses can inform him about common misconceptions and misunderstood points and about points which got across well.

The instructor then need only revise those points which students fail to understand.

Evaluation of a class or classes of instruments

As has been pointed out repeatedly (e.g., see Hovland, Lumsdaine, and Sheffield, 1949, p. 5), conclusions from an evaluative study of a single instrument apply only to that particular instrument, and the generalization of the results to other instruments of the media it represents have, at most, the status of untested hypotheses. Many studies based on measuring the effects of a single instrument, such as a film, have unfortunately had the purpose of wider generalizability rather than the limited technological purpose of evaluating a single product but nevertheless have conformed to the pattern for the more specific technological studies and have had the same limitations on generalizability. In the extreme case, it should be obvious that the effects of a single film cannot be used to "determine the utility of films as educational devices."

Roshal (1960) has delineated two broad classes of studies which, it is contended, have dominated all too much of the history of research on instructional films. These involve evaluative studies in which the permissible conclusions are really limited to the evaluation of the particular film or set of films studied, but which were conducted for the broader purpose of evaluating the capabilities of films in general. Roshal's characterization may be generalized to other instructional media, including instructional television presentations, which comprise loosely defined, broad classes of instructional instruments or materials. Often these have been grouped together as particular media on the basis of some class property which has historical significance, or which characterizes a production or distribution process but which does not define their specific instructional-stimulus properties.

As in the case of experiments on singel instruments, these studies are often "controlled experiments" in that the paraphernalia of control groups, pre- and post-measurements, and the like are introduced. Sometimes the methodology of measurement or experimental design is impeccable, ingenious or both. Whether this was the case in a given experimental demonstration *obviously* affects the tenability of the direct conclusion that the particular instructional instrument tested did or did not in fact produce the learning ~~[illegible]~~ claimed by the experimenter. But whether this was true or not, the experimental results contribute little, as such, in adding to our knowledge of the specific instructional factors that may have been responsible for the observed effect, and thus, per se, they add next to nothing to a science of instruction.

One class of evaluative experiments identified above seeks to compare the comparative merits of two media or classes of media, or to compare ~~[illegible]~~ a given medium or device with something loosely defined as "conventional methods." The latter has usually meant some unspecified combination of presentation by some instructor, plus perhaps some unspecified use of some text or other study material. The crucial weakness of such experiments has been characterized by Roshal (1960 ). In the case of comparisons involving the instructional film, he points out that

> ". . . it was some film, not the best possible film or a representative film (. . . in the sampling sense) but some film compared with some other instruction--not the best possible teacher or a representative teacher, but some teacher" and that ". . . neither the film nor the compared medium is specified analytically or chosen by a sampling technique which lends itself to determination of an estimate of error," so that little or nothing can be concluded that permits any scientifically useful generalization to be drawn.

A similar type of study, most often conducted when a ~~[illegible]~~ device or medium is newly introduced, may be called the evaluative demonstration experiment. It asks, in essence, the questions, Can films teach? or Can television teach? or Can device X teach? More specifically it may go on to inquire, Can medium X (or device type Y) be used to attain a specific class of instructional outcome? Thus, can X teach motor skills, teach factual information, produce changes in opinion, etc.? The status of this kind of study has been described by Hovland, Lumsdaine, and Sheffield (1949, p. 6) in the following terms:

> "For example, a study may be done to determine the effectiveness of films in teaching a particular subject, such as general science. Even if an adequate sample of existing films of this type were used and compared with an adequate sampling of other instructional devices, the conclusion would apply only to existing films of this type and would not determine how effective such films could be."

Nevertheless, a seemingly endless series of studies falling in one or the other of these two classes has been conducted and reported. Studies documenting that films can teach this or that were reviewed at length by Hoban and van Ormer (1950), who also summarized numerous studies presenting evidence that a particular film was shown or failed to be shown superior to some alternative instruction in terms of some level of confidence. More recently, Allen (1960) has similarly marshaled extensive citations of similar results not only for films, but also for instructional television and other forms of audio-visual presentation.

No attempt will be made here to describe of even to cite most of these experimental demonstrations that "X can teach," that "X can teach A," or that "$X_1$ taught a bit more than $Y_1$" while "$X_2$ taught a bit less than $Y_2$," though "$X_3$ and $Y_3$" were not shown to differ significantly (regardless of whether X was a film, a television lesson, or the use of some other instructional medium or device). In the case of films, television, and related media, the interested reader can refer to the reviews by Allen and by Hoban and van Ormer. A few comparisons of this type have been cited elsewhere by the writer (Lumsdaine, 1960a), who characterized the status of such device-device or media-media comparisons as follows:

> "Comparisons between existing devices, procedures, or media that represent mixed aggregates of characteristics can be useful as a basis for making a specific, practical choice between two existing alternative devices or sets of material. However, it seems apparent that most such 'evaluative' comparisons are not likely to improve our understanding of any one of the several factors

that happen to have been thrown into any one 'pot' (device), nor to give us a valid basis for deciding what characteristics future devices should have.

"The inherent weaknessess of such media-vs-media or device-vs-device comparisons have long been recognized as relatively unproductive in terms of any direct scientific contribution--see Hovland, Lumsdaine, and Sheffield (1949) 120-121 and 179-181; Lumsdaine (1953); esp. pp. 85-90. Little can be contributed to the applied science of human learning by adding to the already large number of so-called 'evaluational' experiments that have been conducted to try to compare 'media' or 'device types.'

"These past experiments have, all too often, been foredoomed to failure (so far as any permanent scientific contribution is concerned) because though sometimes employing painstakingly executed experimental designs, the things compared represented undefined or adventitious combinations of characteristics. It was thus impossible to isolate the specific factors to which any differences obtained could be unambiguously attributed."

Can [redacted] device-vs-device, media-vs-media, or media-vs-"conventional" experiments ever have utility? As a basis for scientific generalizations, no. Their [redacted] utility as a basis for choice between two existing [redacted] alternatives has been indicated above, but the logic and consequences of this basis for decision are exactly the same as if the choice is merely between two specific representations of the same medium--i.e., the question, Is Filmstrip A better than Film B? does not differ in scientific status from the even more obviously ungeneralizable question, Is Smith's film ($A_1$) better than Brown's film ($A_2$)?

Under certain [redacted] circumstances, comparison of two non-analytically defined media may be defended, despite the foregoing, but primarily for purposes other than that of providing scientific generalizations. These purposes may be propagandistic, in the interest of gaining support for more incisive [redacted] research and development, or may represent other heuristic aims, such as [redacted] testing the worth of a general approach before proceeding; or to provide experiences, or hypotheses about specific factors that may be a useful background for planned future experiments; or to establish that a sufficient range of effects can be achieved, to insure that a methodology can profitably be further pursued along similar lines; etc.

However, it can be contended that such media-versus-media experiments are justified only under rare circumstances which could include meeting all five of the following conditions.

First: Either:

a. There is a very confident a priori basis for expecting that a new medium, X, [redacted] will show [redacted] sizable achievement that is grossly superior (not just reliably superior in terms of a statistical significance criterion) to some other medium, Y (or to some aggregate of "conventional" procedures); or

b. There is a firm basis for expectation, based on prior estimates of the potency of X and Y and of the sensitivity of the measuring instruments that, [redacted] where X is considerably and consistently cheaper than Y, X will show a somewhat superior achievement to Y at a significant confidence level; or

c. X and Y have a considerable and consistent difference in cost, and there is some [redacted] defensible basis for stipulating that a difference of less than a specified amount in a standardized criterion measure is of less importance than the difference in cost.[36.1]

Second: Either:

a. X and Y are both represented by a reasonable sampling of instances $X_1$, $X_2$, . . . $X_N$ from a tolerably well-defined population class; or

b. Y is thus represented in sampling terms, and X is defined analytically so as to specify its properties, and/or the principles it is desinged to embody, in generalizable form. (A necessary [redacted] condition for this is that, were X and Y both so defined, comparison of X and Y would constitute a scientific experiment to test a hypothesis concerning a single, specified characteristic with respect to which they differed.)

Third. The status of the experimental comparison as an evaluative comparison is clearly recognized and its lack of generality is explicitly stated.

Fourth. The propagandistic or other heuristic purpose of the experiment could not be served as well or better by (1) a less expensive method or (2) by some alternative (such as rational argument) less likely to invite spurious generalization of the results, or (3) within feasible limitations of resources, or by an analytic experiment capable of yielding generalizable findings about the operation of specific variables.[37.1]

Fifth. No similar experiment involving substantially the same media comparison (e.g., television versus live instruction) has been performed which is substantially the same in purpose, concept, general method, and class of training outcome, despite differences in specific subject matter, branch of government, or other institution conducting the experiment, or other such adventitious factors. Application of this condition would, for example, counterindicate all or nearly all of the current "television-can-teach" experiments after the first one.

It can be recommended that the conduct of evaluative media-media comparisons not fulfilling all five of these conditions be largely, if not wholly, discontinued in favor of experiments employing controlled variation of specific factors (see below). Even if there are other extenuating circumstances under which such media comparisons might be warrented, this proscription would probably do much to improve research in the study of instructional media. Much, indeed, could have been accomplished had such a proscription been enforced during the decade 1949-1959.

Because of the restrictions on generality that apply to evaluative studies, only two studies dealing with the newer, self-instructional media will be summarized. The first of these (Porter, 1959) was concerned with teaching spelling to elementary school children. Porter used a simple write-in type of teaching machine (see Lumsdaine and Glaser, 1960). The materials were printed as a series of frames on paper fed through the machine, and required the student to respond by writing directly on the sheet through an open window in the machine. The student then moved a lever which moved the paper up one line; this placed the student's written answer under a sheet of transparent plate where he could no longer change it and also exposed the correct answer so that he could check the correctness of his answer before proceeding to the next item. Teaching materials paralleled, as closely as possible, standard lessons in textbooks used by a conventionally taught control group. The words taught were the same as those taught to the control group.

Spelling programs (see Porter, 1959, pp. 86-87) were used by sixth- and second-grade pupils for about twenty weeks of a normal academic year's work in spelling. Results, measured by standardized achievement tests, showed the experimental group to be "significantly superior" to the control groups at both the second- and sixth-grade levels. For the sixth-grade level, time required for the machine program was reported to be only about a third of the time required for the usual class instruction. Porter's report, however, gives no basis for determining how typical the "conventional" instruction may have been, nor does it indicate whether the significance tests reported took class-to-class variability into account (see page 149-154).

Klaus and Lumsdaine (1960) were concerned with the effectiveness of applying self-instructional techniques as supplements to ongoing instruction. Their specific objective was to determine the contribution made to achievement in high school physics by providing self-tutoring materials in the form of programmed textbooks. The self-tutoring materials covered approximately six weeks of instruction in the second semester of high school physics and comprised approximately 3,000 individual question-and-answer "frames." Results were appraised with comprehensive examinations administered to approximately 450 students in 15 schools in western Pennsylvania which regularly utilized a filmed physics course telecast daily over station WQED in Pittsburgh. These tests were constructed by an independent agency and were not seen by the program writers until after the programs were completed.

Comparison of results for experimental classes who received the self-instructional materials to supplement the physics course (as compared with control classes who did not receive the ~~[illegible]~~ materials) demonstrated that the self-tutoring materials made a significant contribution to the level of achievement of the students. The units of sampling and corresponding units of statistical analysis were intact class groups.

## Controlled Variation of Specific Factors in Design and Use of Instructional Media

### The study of specific factors

If the arguments of the preceding section are accepted, it becomes clearer what kinds of experiments are most needed to construct a usable science of instructional media. What is needed are experiments which seek to reveal the influence of specific factors in the design characteristics of the media. These factors should define reproducible stimulus and response characteristics that can be implemented in future instructional materials and devices. In this way we can obtain experimental data to support the validity of generalizations on which to base future decisions about media design.

It can well be argued also that the precision and applicability of experimental findings are likely to be augmented to the extent that the specific factors chosen for manipulation are defined in terms of theoretically oriented variables, and not solely in terms of gross physical characteristics of instructional media. As pointed out by Hovland, Lumsdaine, and Sheffield (1949, p. 8), definition in terms of theoretical variables seems likely to afford a better base for generalization of results and is also likely to identify variables which have a consistent influence on the outcomes of instruction.

Emphasis on manipulation of single, defined factors in experimentation is not meant to minimize the importance of interactions or cumulative effects of several variables: quite the contrary (see below). But such interaction effects can be established only through analytic experiments in which each specific factor is independently covaried, not through over-all comparisons of an aggregate of factors.

It is, of course, true that it is not always easy to recognize which variables should be regarded as fundamental or primary. Thus, results of an experiment on the effects of a single procedural element may have to be refined by further experimental analysis to isolate the primary factors at work. This will be illustrated presently. It should also be recognized that in some instances there may be a tactical advantage in confounding several theoretically related factors in an initial experiment, in order to maximize the chances of establishing a palpable effect on instructional outcomes, before proceeding with further attempts to isolate more specific primary variables. This tactic might insure a favorable experimental setting for more refined analysis, including adequate sensitivity of measurement to make more precise experimental definition worth while.

Studying the influence of specific factors in instruction by controlled variation should be sharply contrasted with attempts to study such influences by comparing classes of instruments. For example, if the question were: "Which is more effective for a particular educational purpose, color film or black-and-white film?", the specific factor or variable would be the presence or absence of color in the film. Such a question could conceivably be attacked by collecting examples of color and black and white film, trying to assure that they are "comparable" in other respects, and comparing one group of films with the other to determine their "relative effectiveness." As stated by Hovland, Lumsdaine, and Sheffield (1949, p. 6), "The results of the mode of attack have doubtful generality. At best, they could only give the typical effects of the variable as usually employed; when the sampling of films or other instruments is small, as is usually the case, even this conclusion cannot be drawn."

The effects of a particular variable should be studied by manipulating the variable experimentally by controlled variation. This requires constructing or obtaining alternative versions of a particular instrument, which differ only with respect to the specific factor under investigation. Two general classes of such experiments can be distinguished--experimental investigation of a single variable by controlled variation, and experimental analysis of two or more variables in combination.

## Variation of a single factor

In this approach, all factors are held constant except the one being investigated. This requires the use of two presentations, e.g., films, which have exactly the same content but differ with respect to the specified variable. An example is the use of a color film and an achromatic print of the same film to investigate the effect of color. Here, the experimental and control forms of the instrument are constructed so as to be comparable with respect to all factors except one. In the case of many variables, however, it may be difficult to achieve comparability with respect to all variables except one, because differences in content are likely to be associated with differences in technique of presentation.

Even when such difficulties are overcome, a major problem arises in drawing useful conclusions from studies in which only one factor is varied, because of the difficulty of assuring an adequate basis for generalization of the findings. Inherently, experiments which manipulate only a single variable without respect to other variables fail to take account of the possible interaction of the experimental variable and other factors which might influence its operation. For example, the use of color might be an aid to learning under some conditions and not under others, or the effectiveness of review procedures might depend on the characteristics of the original presentation. An unqualified conclusion drawn from a single variable study thus ordinarily needs to be checked with a variety of instruments and under a variety of conditions before its generality can be determined.

Experimental analysis of two or more variables in combination

The case for multivariable experimentation is stated by Hovland, Lumsdaine, and Sheffield (1949, p. 8) as follows:

> "As suggested above, it seems likely that with the complexity of variables present there would be few empirical generalizations that would hold up for all educational films, all audiences, and all conditions for using the films. Variables would be expected to interact so that the effects of any one variable would have to be differentially designated according to the accompanying variables. Accordingly, the result of an attempt to determine the generality of a conclusion about a single variable would lead to a series of principles rather than a single principle.
>
> "Because of this likelihood, the type of research that will probably result in the broadest generalizations for the field of educational films and related media is research studying the controlled variation of several variables in combination. The qualifications on the generalizations are thus

determined, and generalisations may be stated in the form: 'Under condition A, result 1 is obtained, whereas under condition B, result 2 is obtained.'"

A number of the experiments subsequently discussed involve simultaneous variation of two or more factors to afford a basis for this kind of contingent generalisation. As noted by Lumsdaine (1953), however, such variation need not involve highly complex experimental designs. One of the simplest designs--the so-called "2 x 2" factorial design--is especially attractive because it provides an economical basis for observing the effect of a given variable under each of two conditions with respect to another variable.

## Simple versus complex experimental designs

Despite the above-noted emphasis on multivariable experiments, the research methods discussed in this chapter are, in general, relatively simple ones. Although several multigroup experiments are described, highly complex designs have, for the most part, been avoided. This reflects, in part, a prejudice of the writer that, in general, the simpler, more straighforward designs are better calculated, in the present state of the art, to reveal important causal relationships. A carefully formulated problem--though it may deal with different effects of media variables as a function of other variables (such as differences in the intelligence of the learner or in the conditions of use)--generally leads to a relatively simple design affording a few specific comparisons. The author feels that such experiments have generally been more productive than "shotgun" multifactorial experiments that are less sharply focused on specific hypotheses.

Similarly, highly complex analyses, including analysis of covariance, are [redacted] not stressed here. Granted the necessity of employing [redacted] these refined and laborious methods on occasion, their use may often be [redacted] occasioned by the fact that the experiments to which [redacted] they are applied were not cleanly designed or properly conducted in the first place. [redacted] Frequently, more attention to the simpler aspects of design and, in particular, to proper randomization in the design of an experiment will obviate the need for highly refined methods of statistical analysis. [redacted]

In any event, the primary purpose of this chapter is [redacted] to contribute to a sound orientation toward fundamentals of research technique in the experimental study of instructional media. [redacted] This purpose may be better served if the methods discussed are [redacted] not obscured by complex analyses which most students will have to accept without genuine [redacted] understanding. [redacted] (For an amusing commentary on this point by an eminent statistician see [redacted] Kendall, [redacted] 1959) [redacted]

"Hiawatha, nothing daunted,
Called for pen and called for paper
Did analyses of variance
Finally produced the figures
Showing beyond peradventure
Everybody else was biased
And the variance components
Did not differ from each other
Or from Hiawatha's . . . . .
Still, they didn't understand it
So they couldn't raise objections
This is what so often happens
With analyses of variance)."

~

## Simple Analyses and Tests

The analysis of variance is, of course, one of the most powerful tools of statistical analysis, and we do not mean to deprecate its value. However, as suggested by Kendall's poem, analysis of variance is sufficiently complex that it is frequently misunderstood and misapplied. One reason for this is that the computational steps involved are sufficiently involved so that Lavoisier's admonition, quoted by Fisher (of keeping the links between the basic observations and data so short that they can readily be perceived at all times) is difficult, if not impossible to observe. Hence much of the use of analysis of variance, with the general exception of that by professionally competent mathematical statisticians, is generally obscured behind a maze of "works" within a "black box", with the ordinary investigator dealing only with the inputs and outputs of the "box". He may see some of the inner wheels turning around, as like a partially transparent case for a clock, and even see some of the manipulations take place in response to the statistical operations he performs or has his clerk perform; but with the increasing use of computers, these operations become even further removed, so that he has little idea what is going on, other than that he collects some data, and eventually out come some results from the computer.

This is not an argument against the use of computers nor against the use of analysis of variance. But, the point is most seriously made here that, there is generally an intimate dependence of the validity of analyses on the ability to keep close tabs on the relationship between original observations through all stages of analysis, to conclusions.

With complex and computerized analyses, this ability is frequently strained and errors are perpetrated even by relatively sophisticated investigators in using these more complex analyses. It is felt strongly that, wherever possible, a simpler analysis should be resorted to--or at least used as a check.

This notion of keeping the data in view applies at a much more elementary level than analysis of variance. If we take the familiar product-moment corelation coefficient, familiar to all students of elementary statistics, a frequently not heeded admonition would avoid much mischief and also sometimes much labor. In this case the graphing out of a scatter plot showing variable Y plotted against variable X, either for the entire batch of data (if reasonably small) or for a representative random sample of it, if extensive, serves two important purposes. First, for many purposes, it quickly tells the investigator whether the correlation is sizeable enough that it is likely to be worth computing; and, further, it enables the relatively practiced eye to estimate the correlation rather closely without any laborious computations at a glance. Obviously, this is not true for all purposes--e.g., where exact values of correlations need to be obtained.

It is likewise true that in some instances a computerized procedure for performing a number of analyses of variance may be very much worth while. Even in these instances, however, it is urged that some simpler and preferably directly visable way of looking at the data be used--perhaps later, as a check, but preferably first. An example is given in the record experiment reported in Chapter 4, where certain results are quickly apparent from merely plotting the original data, although the investigator may, pro forma or to determine a more satisfactory significance level, nevertheless wish to go ahead and compute a more refined measure of statistical significance (Cf. also Ch. 8,10).

## Formal and informal experimentation

Formal experimentation to test hypotheses has sometimes been subject to considerable criticism from proponents of more emphasis on unstructured exploratory research. However, the formal experiment set up specifically for the test of a particular hypothesis is by far the most dependable way we have at our disposal (if, indeed, not the only way) for putting hypotheses about instructional variables or processes to the test. Only by such experiments can we rule out alternative explanations of observed effects (including, of course, the alternative "null" explanation that they are explainable on the basis of chance factors).

## The Rationale of Experimentation

Gilbert (1960) has recently taken an extreme position with respect to experimentation, which would imply that the use of student-response data as a basis for revision for a specific program may be the only defensible use of empirical data in the study of learning. Gilbert decries the conduct of systematically controlled experiments to test particular hypotheses and instead feels the laboratory should be the place of informal, developmental "try-out and discovery." In connection with the latter, he admonishes us to "resist the temptation to design formal experiments." This is based on the supposition that "you don't want to know whether one method teaches better than another; you want to know what method teaches best."

Although the way in which one discovers the latter without the former is not clear in terms of the requirements of formal logic, Gilbert does make the valid point that a revision method of teaching can display itself in a controlled experimental study only if the really efficient method of teaching is used in the experiment. But this contention also leads us to consider the question of the utility of applied development versus basic research (defined as the testing of potentially generalizable propositions) and, perhaps more cogently, the question of the order in which research and development should proceed. A case can certainly be made for arguing that until one has evolved a quite effective method of teaching by informal trial and error in a particular situation, comparison of specific factors may be conducted at such a low level of over-all efficiency that the influence of variables manipulated may be distorted. An interesting example *within* the fild of systematic experimentation is found in the study by Maccoby Michael and Levine (1961) following up the earlier study of Michael and Maccoby (1953), in which it was found that, in order to reveal motivational effects of active student response (as distinct from direct practice effects), it was necessary to operate at a very low level of over-all motivation.

Even if we dissent from Gilbert's seemingly serious abandonment of the scientific method as a way of arriving at valid inductive generalizations about the operation of specific factors, it must be admitted that the optimum strategy in improving instruction is often far from clear as regards the balance and timing of effort as between informal, "try-out" experimenting and more formal experimentation. We can agree that observations deriving from the former, even if based on only one or two students, may be helpful as a practical basis for revision of a particular instructional sequence without sufficing as acceptable proof of inductive generalizations useful in the future. Perhaps also it will be agreed by many persons that both informal, practically-oriented try-out and the conduct of more formal experiments have an important role to play in the combined research and development effort that is needed to develop a real instructional technology.

A related concept is that of exploratory research, the relation of which to more formal experimentation has been characterized by Lumsdaine (1962) as follows:

"The term "Exploratory Research" is applied to the case where an investigator has a bright idea about a new technique, instrument, or theory, and wants to try it out informally in a variety of ways and see what it leads to, prior to formal experiments to test a specific hypothesis. "Exploratory" laboratory research is sometimes held to be a unique locus for the generation of new hypotheses and of ideas for new instructional techniques. Such inventions may find a more congenial atmosphere for development in exploratory research that is not fully committed in advance to the testing of pre-formulated hypotheses. However, they also arise, not uncommonly, as by-products based on informal observations made in the course of conducting more formal experiments. Thus, as an alternative to purely exploratory research, a good possibility is to provide for exploratory follow-up of hypotheses and new techniques suggested in the course of conducting more structured projects."

## Safe Versus Risky Experiments

There are many experiments that one might consider it desirable to do. An argument for doing even the experiment whose predicted outcome seems obvious is that the history of science is studded with instances in which the "obvious", when put to the test, failed to eventuate. "Obvious to whom" may make a difference, of course. What was obvious to Galileo before he dropped the unequal weights from the tower was the opposite of what was, previously, obvious to the skeptics. Thus one basis for doing an experiment is when there is controversy, or when we have conflicting predictions.

On the side of avoiding the obvious-outcome experiment in favor of the less obvious are considerations of economy. We can not do all conceivable experiments even if we were to spend hundreds of years at it. At least one distinguished experimenter in the field of communications was alleged to try to predict experimental outcomes on the basis of theory and, if he could do so, not to do the experiment.

Putting the controversial question to the test probably brooks, in educational research, a hazard that is more troublesome than the experiment by which the experimenter seeks to verify the obvious for himself. This is the dilemma that arises when the experimental outcome doesn't support what seemed like an ironclad prediction. If all experiments in education were as adequately designed to provide a crucial test as are the best ones in physics, this dilemma would arise less often; the unexpected result would properly be construed as serious grounds for questionning the basis on which the prediction was made. But experimenters do make mistakes (even in physics) and, as has been seen, there are numerous reefs on which the bark of experimental inquiry can, and all too often does, run afoul in the sea of educational research. The fact that it is hard to be assured that measures are properly sensitive and valid, that all relevant variables have been randomized and properly reflected in error estimates, etc., properly invites caution not only to "negative" results (with their inevitable lack of conclusiveness in relation to the null-hypothesis paradigm) but, also, of positive results in most educational experiments. Even replicative verification is not an absolute guarantee, as the same procedural shortcomings can be and sometimes are repeated in successive replications. It is thus not wholly in jest that some sophisticated and quite objective experimenters will maintain, as at least a serious hypothesis deserving investigation, that "when the facts contradict the theory, the facts are likely to be wrong." A logical justification for what might here seem to be a cynical or ascientific dictum is readily seen in the common sense habit of re-checking observations that clearly "don't make sense."

### New role of evaluative experiments in suggesting hypotheses

Despite the foregoing caveats, the measurement methodology and some aspects of experimental design employed in evaluative studies are often useful in studies which employ controlled variation of specific factors. In addition, the analysis of effects on individual points of content associated with particular features of presentation in the instrument being studied can frequently provide leads concerning the factors at work. Sometimes, indeed, using different techniques several times within a film for specific subject-matter topics provides comparisons, relating size of effect to technique employed, which may represent the best evidence available on a particular kind of presentation.

Examples are found in the detailed evaluative studies by May and Lumsdaine (1958) of science films on osmosis and the causes of seasonal change, and in the orientation-film evaluation studies by Hovland, Lumsdaine, and Sheffield (1949).

In these cases the results were not limited to over-all assessment, but included analyses of the specific effects associated with various points of content within the films. Comparisons between large and small effects generated for different aspects of the subject matter by contrasting techniques are correlational in the sense that they only permit statements of the form, "That which was presented by technique X yielded larger measured effects than that which was presented by technique Y." However, such statements, though correlational in nature, should be distinguished from correlational findings of the "those who" type, previously mentioned, which only show a relationship between the effects produced on different groups of persons. The "that which" statement at least involves an association between contrasting techniques, whereas the "those who" statement only shows an association of outcomes which may have been due entirely to predisposing factors.

Although correlational associations relating technique to size of effects are obviously less desirable than controlled manipulation of content, they are fairly efficient as a source of leads about the value of alternative techniques. This is particularly true when contrasting findings about size of effect are found within a single experiment on the effects of a particular instrument. Here the audience is constant, and at least it cannot be disputed that the difference in size of effects is associated with some difference in presentation technique or aspect of subject matter. Even though the presentation technique and subject-matter differences are, strictly speaking, confounded, gross differences in effects can suggest plausible hypotheses about the effects of contrasting presentation techniques. (See Hovland, Lumsdaine, and Sheffield, 1949, Chs. 2 and 3 for examples.)

It should also be noted that in some instances it is possible to produce minor experimental variations on several different points within a study originally designed for evaluation. Thus, within a single instrument involving a range of subject-matter items which can be regarded as substantially independent, one can conduct several "little experiments" by varying the technique of presentation of specific aspects of subject matter. Sometimes, also, it is possible to make "experimental" comparisons where the difference lies not in the instructional materials but rather in the test used. An example is found in May and Lumsdaine (1958, p. 252), where inferences are drawn concerning the degree of transfer resulting from a particular kind of presentation. The inferences are based on the use of alternative test items that required different degrees of transfer from the kind of depiction used in the instructional instrument.

Analyses of this type may considerably extend the usefulness of data from evaluation studies, and thus might seem to counterindicate the proscription proposed above. However, such analyses can also be conducted using data from experiments of the controlled-variation type, in which, as argued below, the over-all comparisons provided also make potentially useful contributions to a more rigorously documented body of generalizable principles. The fact that something valuable can be salvaged from the data of evaluative studies thus should not necessarily imply that they be conducted, but only indicates the potential value of the data they may sometimes yield when the data can be analyzed for purposes other than over-all comparison.

# Chapter 13

# The Research Literature on Educational-Media Experiments and on Student Response as a Factor in Instruction

## Purpose of Literature Review

In this chapter and the next one we shall survey a good deal of the experimental research literature on instructional media.* This review is largely limited to "analytic" or scientifically-oriented experiments in which evidence on the effects of specific factors has been sought (see Chapter 1, p. 32 and Chapter 12, p. 8 ), although a few studies of an evaluative sort are mentioned when they illustrate interesting aspects of methodology. The review is not exhaustive, even for experiments on specific factors, but is intended to be extensive enough to serve its two major purposes, characterized below.

*State of present knowledge of instructional-media variables.* The first purpose of the review relates to content rather than method of research. It is this major objective which has largely governed the organization of the present chapter and of Chapter 14. The primary intent here is not so much to survey the "state of the art" of research methodology as to survey the "state of the science," in terms of what variables have been investigated and what can be concluded, at least tentatively, from the experimentation thus far conducted. It is hoped that this presentation will afford a basis for suggesting some of the major problems to which further research needs to be addressed, by delineating where we now seem to stand in our knowledge of factors governing instructional-media effects. In other words, we are trying to provide a survey of the instructional-media terrain as thus far explored, seeking thereby to identify plateaus of knowledge and valleys of ignorance which represent the points of departure from which further inquiries might usefully proceed. At a minimum, we will try to provide a guide to some central areas around which much of the methodologically more or less acceptable past research appears to have clustered.

Illustrations of methodology. The second purpose of these chapters is to provide a context in which reference can be made, particularly in Chapter 15, to concrete instances of methodological points which have been discussed in more general terms in earlier chapters. To this end, various methodological points are identified by the footnotes and supplementary notes to this chapter and Chapter 14, some of which are then amplified to varying degrees in Chapter 15. These methodological comments may be of value in providing instances of studies to illustrate some of the aspects of method which govern validity, sensitiveness, and other features of design with which we have dealt. However, no attempt will be made to do more than comment selectively on methodological points; a complete methodological analysis of all of the studies reviewed in the present chapter and in Chapter 14 would require much more space than can be devoted in the present volume.

## Descriptive categories for reviewing research on instructional media

Previous reviews have commonly been organized in terms of such "media" categories as motion pictures, filmstrips, television, recordings, teaching-machine programs, and the like. However, we believe that this organization is much less likely to reveal meaningful communalities, and thus potentially to contribute to a science of instruction, than is an organization in terms of the independent variables (or stimulus features) in instructional sequences, regardless of the mechanics of presentation for a given sequence. Accordingly, the review in this chapter and Chapter 14 is organized in terms of particular stimulus and response features of instructional sequences, regardless of which media are used to exemplify or implement these features. For the most part, we shall be concerned with the specific physical media used in producing or disseminating instructional material, only as these have an explicit bearing on describable stimulus and response variables.

Some of the categories of stimulus and response variables to be considered include techniques for elicitation of appropriate learner responses, methods for guiding and prompting these responses, knowledge of results, feedback and reinforcement, the organization and content of practice and review, and factors in the sequencing and pacing of instruction. Organization in terms of psychological categories such as these rather than in terms of media categories will facilitate our introduction to relevant "pure" research on learning conducted in psychological laboratories. A more important advantage of utilizing such psychologically defined categories instead of classifying studies by media is that findings having potential generalizability are brought together regardless of their source in terms of the particular media utilized in obtaining them.

## Specific Educational Media on which Research has been Conducted

Since the review of the experimental literature in this chapter and in Chapter 14 is organized in terms of psychologically or pedagogically defined variables affecting learning (cf. Chapter 12) rather than in terms of media, a brief overview of the research which has been conducted using the major classes of media will be presented first in order to provide perspective and relate the present review to the more common media-oriented discussions.

Experiments on instructional films

Up until the past two or three years at least, by far the most extensive experimental research on instructional media has been conducted with motion pictures, and their close relative, the sound film strip. These two media have until recently been the ones most amenable to research because of their dual properties of flexible stimulus control and of reproducibility. However, with the advent of widespread interest in teaching machines and individual programmed instruction, this picture is rapidly changing, and one may anticipate an increasing concentration of research on instructional variables centered around individual self-instruction programs during the decade of the sixties. Films and filmstrips present a sequenced, pre-paced presentation package, complete with visual material (still or moving) and accompanying sound (consisting of either dialogue, narration, music, and sound effects, or some combination of these). In terms of range and variety of controllable stimulus input, these media still have unrivaled advantages for experimentation on instructional variables, despite their lack, as conventionally employed, of the special advantages offered by individually-paced instruction (see below).

The history of experimental research on silent and sound films covers a period over four decades, beginning during World War I. Most of the earlier studies, in the 1920's and 1930's, were directed toward evaluation of the effects of particular films as indicative of the educational capabilities of motion pictures. Among the first of the more analytic experiments on specific factors to appear were the World War II studies of Hovland, Lumsdaine, and Sheffield (1949). The interaction between psychology and education in the development of instructional science and technology (see Melton, 1959; Lumsdaine and Glaser, 1960, pp. 94-98 and 563-572) is reflected in the fact that, although most of the research conducted during the twenties and thirties was carried out by individuals within the field of education, one of the earliest major studies on the effects of instructional media (Lashley and Watson, 1922) was the work of two psychologists. Research workers with backgrounds in experimental psychology have also been largely responsible for the increasing number of analytic experimental studies of film (and--more recently--individual programmed instruction) conducted during the past two decades.

The research on instructional films from 1918 to 1950 was summarized in a volume by Hoban and Van Ormer (1950) published under a Navy-sponsored instructional film research program at Pennsylvania State University. (See also Lumsdaine's review of this volume in the Audio-Visual Communication Review, 1953, 1, 176-184.) Roughly two-thirds of the Hoban-Van Ormer book was devoted to purely evaluative studies and only about one-third to analytic studies of specific factors. Later reviews (e.g., Allen, 1960; Silverman, 1960) have concentrated more heavily, though not exclusively, on analytic studies.

Several major programs of intensive research on the characteristics of instructional films have been carried out since 1940. Four such programs should be mentioned here (see also the information given below on the publications resulting from these programs). The first of these programs was the series of studies conducted in connection with the Army's Information and Education program during World War II and subsequently reported in the book by Hovland, Lumsdaine, and Sheffield (1949). A second sizable program of research was conducted at Yale University from 1946 to 1949, and was subsequently reported in the book by May and Lumsdaine (1958). These two volumes should be considered as standard reference sources for instructional-media experimenters. The most extensive single program of experimentation dealing with instructional films was the program conducted under the sponsorship of the Office of Naval Research at Pennsylvania State University, under the direction of C. R. Carpenter, between 1947 and 1954. This work, which dealt with a wide range of film topics, experimental variables, and types of experimental design, has been reported in a series of more than 50 individual mimeographed reports issued by the U. S. Naval Training Device Center,* which have also been bound together to form two large volumes of reports (see below). A fourth major program of research was conducted through an in-service and contract program of the U. S. Air Force, under the direction of A. A. Lumsdaine, from 1950 to 1953, and was continued, less intensively, in research on other forms of training aids and devices, from 1954 to 1957. Reports on this work, aside from a few journal articles and papers presented at professional meetings, have until recently been largely confined to mimeographed Air Force reports with quite limited circulation. (See Lumsdaine and Glaser, 1960, pp. 574-579, in regard to availability of these and related government reports.) However, a central group of these Air Force studies, dealing with student-response procedures used in conjunction with instructional films, has been recently published under sponsorship of the National Academy of Sciences--National Research Council (Lumsdaine, 1961d). This group of studies (see below) includes a large majority of the experiments on student-response variables conducted prior to 1960, and thus is also regarded as a major source of primary reference for instructional-media experimenters.

Abstracts and summaries of a number of the studies conducted under the Navy Instructional Film Research Program at Pennsylvania State University and under the Air Force program have appeared in the *Audio-Visual Communication Review* and in *Contemporary Psychology*. Some of the findings have also been discussed in other works, of which the three principal ones are: *Graphic Communication and the Crisis in Education* by N. E. Miller and collaborators (1957), an excellent but not widely available summary by Cook (1960), and the readily accessible, less theoretically structured summaries by Allen (1956, 1960). The summary by Cook and the second part of the report by Miller and his associates deal systematically with specific variables and are both organized in terms of factors grouped under the four headings of "Drive," "Cue," "Response," and "Reward." The Allen summary employs the major headings of "Effectiveness of Audio-Visual Materials" (evaluative studies of motion pictures, television, and other audio-visual media), "Audience-Learner Characteristics," "Characteristics of the Learning Environment," and "Instructional Use of Audio-Visual Materials," with most of the studies of specific variables summarized under the first, second, and fourth of these topics.

## Sound Recordings

Relatively little experimental research on instructional variables has been conducted using purely audio presentation without visual accompaniment, although a considerable number of research studies on audio-visual media have used transcriptions or tape recordings to accompany a sequence of visual materials presented by film strip, by a series of slides, or occasionally by accompanying booklet materials--e.g., Kimble and Wulff (1953; 1961 a,b), Newman and Highland (1956). Somewhat more extensive use has been made of audio materials in studying the variables underlying the effectiveness of persuasive communications--for example, the reports by Hovland, Lumsdaine, and Sheffield (1949, Chapter 8), Hovland, Janis, and Kelley (1953), Lumsdaine and Janis (1953) and Hovland *et al*. (1957). An interesting series of experiments dealing with utilization of electronic speech compression techniques was reported by Fairbanks, Guttman, and Miron (1957 a,b,c). These attempted to improve the effectiveness of audio-recorded instruction by compressing more information into a given period of time, through an electronic means which speeded up the speech without distorting the pitch or other characteristics of the sound. Some pilot experimentation on the use of audio-recordings for on-the-job instruction was reported by Hoehn and Lumsdaine (1958). Also to be noted is the quite extensive literature dealing with recordings for language instruction, including their use in the "language laboratory." (Cf. Oinas, 1960; and Carroll, 1962.) Most of this literature, however, is of a descriptive sort, and little controlled experimentation on specific factors governing the effectiveness of the instruction is to be found as yet.*

Instructional television

The rapidly expanding literature of studies on instructional television is heavily loaded with descriptive articles, including several detailed accounts of "experimental" programs introduced into school systems by closed circuit or open circuit television. In many instances, no valid data on effects are presented to supplement the information given about the content of the lessons, cost factors, etc. However, a considerable number of evaluative experimental studies have also appeared, mostly of the "cross-media," TV versus "conventional-instruction" type (see Chapter 12,

Informative accounts of the development and utilization of techniques for televised instruction in the military services have been presented in a series of articles by Kanner and others (cf. Kanner, 1960). Among the better known reports of the development of televised instruction in public schools and colleges are those of the experiments at Hagerstown, Maryland (Weiss, 1957), Pennsylvania State University (Carpenter and Greenhill, 1955, 1958), San Francisco State College (1958), the Los Angeles City School Districts (1959), and the National Workshop (1959). Recently, a number of analytic experiments on specific factors in instructional presentation by television have been conducted under projects supported by Title VII of the National Defense Act, some of which have investigated the possibilities of television for supplementing conventional classroom procedures. Thus far, most of these studies have been purely evaluative in nature; as yet relatively few experiments of an analytic or scientific nature have used television as a medium. An example of the studies attempting to deal with specific variables is the study by Schwarzwalder

(1961) who assessed various methods of supplementing in TV science lessons and reported differences among treatments, indicating that techniques such as superimposition may add to learning.

Among the studies thus far reported, the most extensive single series of experiments is that reported by Gropper, Lumsdaine, et al. (1961 a,b, c,d,e). These studies have dealt with the importance of (1) methods of lesson preparation and "pre-testing" or tryout, which they demonstrated could profitably utilize student responses as a basis for revising a television lesson; (2) methods of eliciting active student response during instruction, some of which were shown to increase the efficiency of learning; and (3) methods of encouraging student follow-up activity as a manifestation of active interest in the subject matter stimulated by the television presentation. These studies also illustrate several methodological points relative to measurement procedures and to group assignment to experimental treatments.

Programmed individual instruction and teaching machines

As has been mentioned previously (see Chapter 2, p. 22) the control of learner behavior and feedback which is provided by the continuous record of student response from programmed self-teaching materials ("auto-instructional" programs) may afford the most promising single vehicle yet developed for the analytic experimental study of variables affecting human learning (as well as for the incorporation of research findings into improved instruments). The great surge of interest in research (as well as development and evaluation) evidenced during the past few years has led to an increasing number of experiments on programming variables, though the number of well-controlled analytic experiments published on the newer forms of self-instructional programs has thus far not caught up with the volume of experiments conducted on instructional films. A few of these more recent programmed-instruction experiments are summarized in the following pages, along with the experiments on other media.*

The recent experimental research on programmed instruction has been more widely dispersed--that is, less concentrated in a few major programs --than has been true of the large proportion of experiments on variables in instructional films. The original developments of self-instructional devices, and a few experiments on specific factors, originated in Pressey's work at Ohio State University (see reports by Pressey and others printed in Part II of the 1960 Lumsdaine-Glaser source book). The further development of small-step programmed sequences, originated by Skinner and his collaborators (see Parts III and V of the same source book), represented an innovation of very great significance to education. However, Skinner's program concentrated mainly on development of rationale, programs, and devices and, though it has included a few evaluative and other experiments, has not stressed controlled, analytic experiments to assess the influence of specific factors.

Among the closest approaches to these sustained programs of analytic experimentation on self-instruction in which specific factors of programs have perhaps been those under Robert Glaser's direction at the University of Pittsburgh, those by Coulson, Silberman and their colleagues at the System Development Corporation, and those conducted under several U.S.O.E. Title VII grants by Angell, Lumsdaine, Briggs, Goldbeck and others at the American Institute for Research.* However, so many other investigators have begun to report individual experiments or small sets of experiments that the total number of such isolated experiments probably exceeds the number as yet conducted as parts of larger programmatic efforts.

On the other hand, it should be noted that, to a much greater extent than has ever been true in the case of instructional films, evaluative research aimed at product improvement on the basis of tryout of preliminary programs is a very frequent concomitant of the development of autoinstructional materials (and would be a universal one if the position advocated by most workers in this field were fully adopted). A number of examples of such empirically-based development of materials, in which student responses to self-instructional frames were used as a basis for program sequence revision, have been reported. Such studies as those by Meyer (1960), and Klaus and Lumsdaine (1960) have made immediate cumulative revision of program frames based on intensive work with a relatively small number of cases. A different pattern is illustrated by Holland (1960), who, as in the television studies by Gropper and Lumsdaine, obtained responses from a considerably larger group to an already completed program. The study by Klaus and Lumsdaine (1960) illustrates the frequent finding that step-by-step revision can usefully be based on discovered misconceptions in the responses of even a very few students. Such a procedure permits a cumulative revision which, in some instances, might be more useful in product development than a longer-deferred, larger-scale study.*

## Non-programmed instructional materials

As was indicated in Chapter 1, static instructional aids such as graphic aids, charts, the chalkboard, isolated slides, models, and mockups lend themselves to experimental research only if they are considered elements that are integrated, sequenced, and governed by an instructor's presentation, recorded or otherwise. (Several studies of such presentations are mentioned below in relation to specific variables studied experimentally.) When visual aids are used in this fashion, there is some evidence to indicate that their contribution can easily be exaggerated (see Swanson, Lumsdaine, & Aukes, 1956). An analysis of types of mock-ups, models, and other graphic aids used both as self-contained training devices and (with appropriate programming) as aids to lectures has been presented by Lumsdaine (1960a). Several studies attempting to evaluate classes of devices have been made; an example is the study by Torkelson (1954), who endeavored to compare the effectiveness of a mock-up, a cut-away, and projected charts in teaching nomenclature and equipment function, and have not found evidence to support hypothesized differences among these "classes" of visual aids.

The literature dealing with textbooks (see the summary by Buckingham, 1960) contains, as previously indicated, relatively few experimental studies of effect; empirical research in this field is mostly limited to content analyses and readability studies (e.g., Flesch, 1949) which, in almost all cases, have used correlational comparisons rather than experimental manipulation of a particular piece of instructional content. (In contrast, see the experimental study of the "readability" of film commentary reported by Gladstone, 1958.) Perhaps the most extensive series of studies in which factors affecting comprehension were experimentally manipulated was that conducted by Klare and his associates (Klare, et al., 1955 a, b, 1957, 1959). Controlled reading of text presentation has also been compared experimentally, as a "control" baseline, with "programmed" text materials of the self-instructional sort (Goldbeck, 1960; Holland, 1960a).*

## Research on characteristics of instructional equipment

Many instructional materials require capital equipment (in the form of projectors, transcription players, response-registering equipment, and other apparatus) in order to be converted from inert material into functioning instructional media. Experimental research on the characteristics of this associated equipment would thus appear to be a legitimate concern. In some instances, the functional characteristics of the media are intimately bound up with specific features of the presentation equipment, as has been noted previously. With few exceptions, however (e.g., Ash and Jaspen, 1953), relatively little experimental research has been devoted directly to the functional or physical characteristics of equipment. Implications for functional characteristics of start-stop, daylight loading, automatic-threading film projection equipment emerge directly from some of the studies by Maccoby and Sheffield (e.g., 1958) and others on alternate film demonstration and practice, though these studies were not aimed directly at determining projection-equipment characteristics.

There is, of course, an intimate connection between the content of self-instructional programs and the equipment features related to provision of prompts and, particularly, to response mode and automaticity of feedback in the case of teaching machines. Accordingly, an increasing number of studies would appear to have implications for equipment characteristics. Some of the variables which experimenters have sought to investigate include mode of response, use of branching sequences, and various forms of prompting. Several studies of these kinds are reported in the present chapter, though the orientation has generally been as much or more directed toward the programming aspect than directly toward implications for equipment designs as such.

## Treatment of methodology

As noted above, the review of the literature presented in the rest of this chapter and in the next one stresses findings rather than method *per se*. Since the purpose of this review is primarily to identify areas of knowledge and ignorance from which further hypothesization and experiment might proceed, methods whereby we have arrived at the present state of the science are in general unimportant here. An exception, however, lies in some instances where basic aspects of the methods actually help to define the hypothesis investigated or to condition the meaning of the results obtained. But the experimental structure and sampling arrangements are, as such, generally not relevant in this respect, provided only that the design, whatever it was, was a sound one capable of yielding valid conclusions. Structural aspects are therefore minimized in the present account. (An exception to this would be where the findings are, because of structural defects, of questionable or borderline validity; in a few instances of this kind where it seems nevertheless desirable to report a study because, despite its defects, it still represents partial or at least suggestive evidence deemed of some consequence, the findings are suitably qualified and reference is generally made to later discussion where the methodological aspects are further dealt with. [in the supplementary notes to this chapter and Chapter 14, and also Chapter 15]). In a few other cases, moreover, it is convenient to give only an approximate or incomplete account of findings without discussing design features; here the choice is between at least mentioning the necessary methodological aspects in the present chapter or deferring the statement of findings to the following chapter in the context of discussion oriented primarily toward method. In such cases, a cross-reference is generally made.

Special reference designations

The form of reference citation adopted for use in this chapter and in Chapters 14 and 15 has been chosen to assist the reader in locating relevant papers rather than adhering uniformly to the standard form of citation. A great deal of the relevant literature--half or more of it, probably--is contained in a relatively few major source books. For these papers, citation in the text to the most convenient source is given even where this differs from the original, less readily accessible source. However, the form of citation used also makes it easy to identify the original source by consulting the Bibliography (see Appendix A). The form generally used is to indicate author and year (with, where appropriate, designating letter--(a), (b), (c), etc.--within the year) of the original publication (or in some cases the latest version of it, if there is more than one publication), followed by one of the abbreviations listed below to indicate the source book where the publication can be found. (In the following material, to avoid confusing these outside references with chapter references to the present book, chapter numbers of outside references will be abbreviated Ch. and underlined--e.g., MROT, Ch. 10--and chapter references to the present book will be written out, without underlining--e.g., Chapter 3.)*

The following are the source abbreviations used for this purpose. ("L.C." stands for Library of Congress catalog card number.)

HLS = C. I. Hovland, A. A. Lumsdaine, and F. D. Sheffield, Experiments on Mass Communication (Princeton University Press, 1949; LC#537776; $6.00).

**M&L** = M. A. May and A. A. Lumsdaine, Learning from Films (Yale University Press, 1958; LC#58-6541; $5.00).

**TMPL** = A. A. Lumsdaine and R. Glaser (Eds.), Teaching Machines and Programmed Learning: A Source Book (National Education Association, 1960; LC#60-15721; $7.50).

**IFRR** = Instructional Film Research Reports, Vol. I (1952) and Vol. II (1955) (Based on the program at Pennsylvania State University [see above]; published by Naval Training Device Center; available for purchase from the U.S. Dept. of Commerce's Office of Technical Services;*

**HROT** = N. L. Gage (Ed.), Handbook of Research on Teaching (Rand-McNally, 1962;

**SRPI** = A. A. Lumsdaine (Ed.), Student Response in Programmed Instruction (published by the National Academy of Sciences - National Research Council; but available only from the U.S. Department of Commerce's Office of Technical Services; LC #61-60098; $7.00

## Limitations of this review

The importance of the availability of such more complete sources as these for references in studying past research in relation to proposed further experiments is to be emphasised, because of the need for as close a study as possible of what was actually done in order to interpret as fully as possible what was actually found out." While this cannot be fully reflected in the present survey without losing sight of the woods for the trees, the complexities and subtleties of variation in the definition and implementation of both independent and dependent variables in instructional-media experiments is so great, and such standardisation of descriptive terms as exists is at present so poor, that any brief summary of findings must be regarded as no more than a rough first approximation of what has, in fact, been found out. The present survey, as well as others referred to above, is no exception; despite the attempt to be as accurate as is possible in making such summaries, they are to be viewed in the light of a remark once made by E. R. Guthrie to the effect that the information presented in any introductory survey (his specific reference was to introductory courses in psychology) necessarily must ~~be viewed as some~~thing that is "approximately true but yet understandable by sophomores and compressible within the limits of one semester." In like vein, though the present survey is believed to give as accurate an overview as is compressible within the limites of a reasonable-sized chapter, it *is* an overview and necessarily blurs many details, some of which will turn out to be important and indeed essential to consider in the intelligent planning of further research. It is therefore earnestly recommended that detailed study be made, by the prospective instructional-media researcher, of the basic reports in at least those experiments which such a survey as the present one indicates are closely relevant to the particular problem he proposes to investigate."*

Even as a rough overview, no survey of findings such as that presented in these two chapters can ever be complete, since the body of knowledge on instructional media, as in other areas of inquiry, is constantly growing; viability is a characteristic of any science and of any field of inquiry that hopes to become a science (cf. Chapter 5, pp.    ). Since new studies are continually being performed and reported, any such survey as this is thus bound to be partially outdated before it appears in print. Even in surveying past research, we have not attempted to catalog all of the past research that has been done on instructional films and other visual aids, nor to cover comprehensively the very extensive range of related research on conditions that influence the efficiency of learning. Finally, as noted in Chapter 1 (p. 18), the broad area of propaganda, indoctrination, and persuasive-communications research has been largely by-passed in this review in favor of concentrating on studies lying within the somewhat narrower range of instructional objectives comprising factual knowledge and understanding, verbal and symbolic skills, perceptual motor and procedural learning, and, to some extent, interests and motivations. However, we have included in Chapters 14 and 15 a few illustrative studies of persuasive communication which illustrate important aspects of methodology. We have also drawn occasionally from related research in the field of laboratory experimentation on learning which, though not necessarily conceived originally to be concerned directly with practical problems in the design and utilization of instructional media, nevertheless provides instructive inputs to the research area with which this book is primarily concerned.

## Organization of the review

The review of experimental findings on instructional media in the rest of this chapter and Chapter 14 parallels that of the review presented by the senior author (HROT, CH. 12) and also, in abbreviated form, in the ninth chapter of the volume edited by Glaser (1962), Training Research and Education. However, because of the previously-noted emphasis in the present chapter on presenting findings as perspective for the identification of research problems, some of the less conclusive or less interesting results are for brevity omitted here and most of the methodological aspects of the studies are deferred for later, more systematic consideration in Chapter 15.

As in the two parallel reviews (Lumsdaine, HROT, Ch. 12; Lumsdaine, 1962), the results of a particular "study" or "article" have not always been reported in one place within these two chapters, particularly when a study deals with several variables and when the results can be treated more meaningfully in relation to each of two or more particular classes of variables than by being grouped in terms of the historical fact that they happened to be dealt with in a single experiment. For example, some of the results of a study by Michael and Maccoby (1953, SRPI, Ch. 18) are reported in relation to the role of active response, and other results from the same study are reported in relation to knowledge of results and feedback. Similarly, some of the results of a study by Coulson and Silberman (1959, TMPL, pp. 452-468) bear particularly on the question of response mode and are discussed in relation to other studies bearing on this question, while other data from the same study are discussed in relation to the organization and "size of steps" in instructional sequences.

The rest of the present chapter will be concerned primarily with the sizable concentration of experiments which have been oriented in terms of factors directly influencing explicit responses by the student during the course of instruction. As has been briefly indicated in Chapter 1 (pp. 21-22) and argued in more detail elsewhere (see Lumsdaine, SRPI, Ch. 1 and 31), this focus of experimental inquiry is of particular importance in developing an instructional science and technology. It is, furthermore, emphasis on student response factors which has, more than any other single aspect, differentiated recent research and development in programmed instruction from earlier research on educational media. Somewhat paradoxically, however, most of the early experimental work on student-response factors actually stems from research programs concerned with instructional films prior to the more recent surge of interest in teaching machines and the kinds of self-instructional programs they currently are used to present; see below. In any case, the concentration of studies which appear to be most meaningfully described in relation to student-response variables form, to date, so large a proportion of the total volume of scientifically oriented instructional-media experiments that it will comprise vitually all the studies reviewed in this chapter. The following chapter (Chapter 14) will include a somewhat less unitary group of studies dealing with a variaty of stimulus factors and variables in the structuring and organization of instructional sequences.

## Studies of Student-Response Factors in the Design and Use of Instructional Media

We have previously emphasized the importance, for the experimental study of instructional media, of focusing on variables that seem calculated to foster effective control of student response during learning. As mentioned previously, an extensive source of studies pertaining to this class of instructional-media variables is the recent symposium volume, SEPI.

Let us now look at the results and implications of some of these and related studies of student-response factors in instruction. As a point of perspective we may note that both teachers and educational psychologists have long been aware, in a general way, that students may learn more efficiently if provision is made for frequent recitation or rehearsal of appropriate responses that provide practice of what is to be learned. Such active recitation or rehearsal is of both practical and theoretical interest.

### Procedures for fostering explicit student response during instruction

Concern with active,explicit student response as an experimental variable reflects Dewey's emphasis on "learning by doing," which was elaborated in behavioral terms by Smith and Guthrie (1923) and has since occupied a central role in learning theory (and, sometimes, in educational practice). However, unlike some progressive-education proponents, Guthrie, Skinner,and other psychological theorists have stressed not merely "doing" (something) but the controlled elicitation of desired responses in contexts to which instructional objectives require that they become associated.

Experimental studies of the value of active recitation, "participation" exercises, and the like stem from early studies by Witasek (1907) and Gates (1917). The advantage they found for active recitation over mere reading forms a point of departure for a number of subsequent studies in which the function of active response by the student has been studied experimentally. It also represents one of the historical antecedents for current stress on continuous active responding by the student in the use of programmed self-instructional media.

The effectiveness of overt active-response procedures

The potential advantages of using active-response procedures in employing instructional media has been studied in a number of experiments. In the most widely cited of these experiments, performed during World War II (HLS, Ch. 9), a film was used to teach the military phonetic alphabet, Able for A, Baker for B, and so forth. Two forms of the film were compared, which differed only in the review sequences that followed the initial presentation of the letter-word equivalents. The control film used a standard, "passive" form of review, in which letters were presented along with their phonetic equivalents. This was found to be inferior to the active-review film, in which the letters only were presented during the review, during which the audience members were instructed to try to call out the correct equivalents for each letter in turn. The rationale and findings of this basic experiment are summarised in several sources (e.g., HRCT, Ch. 12; Lumsdaine, 1959, Ch. 15). The serious experimenter should, however, read the original account given by HLS (Ch. 9); see also the discussion of methodology for this experiment in Chapter 15. Further analysis of the data by Hovland, Lumsdaine, and Sheffield showed that the difference in favor of the active-response and feedback procedure appeared to be least when least needed and most when most needed--that is, it was greatest for less motivated, slower students in learning the more difficult portions of the material, and least for brighter, highly motivated students in learning the easier portions of the material (cf. Chapter 15).

An experiment by Lumsdaine and Gladstone (MML, Ch. 4) provided a replication of the HLS study, using the same subject matter, but differing in various aspects of experimental procedure and criterion measures employed. (Further results from this study are given later--see Chapter 14). The results on active response replicated the previous findings, indicating superiority for active-response procedure over passive review. In these two experiments the subject matter was verbal in nature, and the task was a matter of simple one-to-one association between paired terms. Such tasks are of considerable prevalence and importance (e.g., learning to pair foreign-language vocabulary terms, names and faces, commodities and prices, objects and terms of nomenclature, etc.). Thus direct application of the method in many practical learning situations is possible, in addition to the more general theoretical interest of the results. (Cf. SRPI, Chs. 20-30 for other studies using paired associates materials.) The use of active-recitation procedures has been studied experimentally for film-based teaching of various other kinds of tasks, including learning of facts and principles (e.g., Michael and Maccoby; SRPI, Chs. 18 and 19), procedural skills such as equipment assembly (e.g., Maccoby and Sheffield; SRPI, Chs. 3-9), intellectual skills such as reading of slide-rule scales (e.g., Kimble and Wulff; SRPI, Chs. 15 and 16). From a considerable stream of experiments following up the study of this kind of procedure, mostly reported or summarized in SRPI (1961), has come considerable data concerning hypotheses ⟶

about specific conditions that influence the effectiveness of active response interspersed in the context of expository or demonstrational instruction. The evidence obtained with respect to several specific variables is summarized below.*

In a 1946 experiment at Yale University (see MHL, Ch. 6), use of pupil participation questions with the correct answer given after pupils had responded was found to add significantly to learning of substantive information learned from a film on the heart and circulation of the blood. It was noted that such devices could be used by producers in making new films or in revising old ones with very little expense to the producers or the owners. This and related studies also suggest that for teaching purposes more liberal use could be made of titles, questions, and other printed words. As noted by the authors, "There is no reason why the verbal aspect of a teaching film should be exclusively auditory." (MHL, Ch. 6.) In this experiment, the "participation" questions were added to the basic film rather than replacing sections of it, and added some 4-1/2 minutes to the film-showing time of 8-1/2 minutes. However, it was noted that showing the film twice, which took about 17 minutes, did not yield appreciably higher scores than the single showing with active-response exercises, which took only 13 minutes. (See the discussion in Chapter 15 on methodological problems related to differences in instructional time.)

In assessing the utility of a procedure such as that of methods for fostering active student reponse, it is clear that not only time expenditure but also other factors entering into the cost of the instruction, including the cost of preparing instructional materials, need to be considered. For example, a quick, low-cost production method might result in less than optimal effectiveness, but thus disadvantages might be offset by the introduction of an active-review procedure. This might or might not require additional instructional time; if so, this factor would of course have to be weighed against the decreased cost of producing the materials. An example is found in an unpublished study by Duva and Lumsdaine (1956), in which adverse effects of relatively poor quality resulting from a quick, kinescope method of reproducing an instructional film were offset by the introduction of student participation procedures. Although the inferior-quality kinescope was less effective than a superior-quality motion picture with parallel content when such procedures were not utilized, this differential disappeared when brief student-participation exercises were added to the lower-quality (kinescope) version.

In some experiments, students have been required to try to carry out active-practice exercises during the showing of a filmed demonstration. Roshal (IFRR, 1949; see SRPL, Ch. 11) used this procedure for some of his experimental subjects while they viewed a film on how to tie knots. Unlike other variables manipulated in this experiment (see Chapter 14), the active-practice procedure did not yield a significant increase in proficiency. The lack of effectiveness in this case was attributed by Roshal to insufficient time for effective practice, carrying with it the further deleterious consequence that students' attention was divided between trying to watch the demonstration and trying to practice tying the knot. Results of a later experiment by Jaspen (IFRR, 1950) lend some support to this explanation. Jaspen studied the influence of several variables on the effectiveness of a procedural film designed to teach his Ss (Navy recruits) how to assemble the breech-block of an anti-aircraft gun. He found that provision for active practice during the film was effective with a slow-paced version of the film which allowed students to perform the actions demonstrated and still be able to attend to the demonstration as it proceeded. However, with a faster-paced film, Jaspen's data, like Roshal's, showed no significant improvement as a consequence of attempted practice during the film showing. These findings (see also the experiment by McGuire described on p. 36 below) are of interest in relation to the results of Ash and Carlton (IFRR, 1951), who found no advantage for note-taking during a film (when notes were collected and tests administered immediately after the showing). With a film deliberately planned to allow enough time for effective note taking (and/or with conditions allowing realistically for using notes for review), it seems not unlikely that a different result would be obtained.

The feasibility of obtaining gains in instructional effectiveness through the use of procedures to provide for active student response has been shown in the case of televised instruction in several recent experiments by Gropper and Lumsdaine (1961; see also summary in HROT, Ch. 12). In one of their experiments, a conventional 25-minute television lesson on physical and chemical changes was prepared for 7th grade students by a qualified science teacher and was found to be less effective than an experimental presentation covering substantially the same material. This experimental presentation provided frequent opportunity for active, explicit responding as the lesson proceeded. (For example, incomplete sentences were shown on the screen, and the students were instructed to fill in the missing word or words, on blanks that had been distributed to them, before the correct answer was flashed on the screen.) The sequencing of the lesson was organized to maximize the likelihood that correct responses could be made at each point.

Other studies on the use of active student responses have been reported by Hirsch ([illegible], 1952) and by Kurtz, Walter, and Brenner (IFRR, 1950). These investigations have employed a variety of questioning procedures and have provided some further information on conditions that influence the effectiveness of such techniques. Despite these positive findings, the advantages of active, explicit student response are not clear-cut in all instances.* Such experiments as those of Michael and Maccoby (1953, SRPI, Chs. 18 and 19) and Cook and Spitzer (1960, TMPL, p.604) failed to find, for some instructional situations, overt responses more effective than "covert" or "mental" responses (see also p.35, below). Also, the 1960 study by Goldbeck found that, for some kinds of material, merely reading passages in which key words were underlined was sometimes more effective than having students make the active, explicit response of filling in the blanks mentally or in writing. (Further results are discussed on p.36, below.)

Relative amounts of presentation and recitation or practice

One of the questions that can be asked about the use of practice exercises or other forms of active response is what proportion of a constant amount of time should be devoted to active, explicit responding (practice, recitation, etc.) versus the amount devoted to exposition demonstration and similar implicit-response activities. This was the original question posed by Gates (1917), and surprisingly little further direct evidence has been obtained subsequently on this question to guide the efforts of instructional programmers and teachers. In comparing the effectiveness of various allocations of total learning time between merely reading the material and attempted recitation, Gates found that, although the amount learned increased progressively with increasing amount of time spent in recitation, the effect was much more marked for learning isolated syllables than for learning prose material (biographies)--a finding that is relevant to subsequent experimentation, summarized below, and that is also relevant to current prevalent interest in programmed self-instructional materials that stress overt responding in the learning of so-called "connected-discourse" subject matter. Woodworth (1938) attributed the relatively slight gains due to active responding in the case of prose material to the fact that, in re-reading such material repeatedly, the student is likely to anticipate what is coming and consequently engages in recitation while reading. Such implicit recitation during the course of reading may occur also, of course, even during a first reading if the material is structured, as in the case of many recent programmed-instruction materials, so that gaps left in the material could readily be filled in, covertly or overtly, by the learner.

A more recent study (Kimble and Wulff, SRPI, Ch. 16) included two experiments in which a combination of booklets and tape recorded instruction was used to present instruction on the reading of slide rule scales. Examples in the booklets were either accompanied by explanation (expository examples) or required the student to determine the answer (participation examples). From the results of the two experiments, the experimenters concluded that the optimal proportion of active-response or participation examples was probably between 50% and 75%. (However, differences significant at a satisfactory level of confidence were obtained in the analyses presented by the experimenters only in comparing 50% and 100% levels of practice, in one experiment, and in comparing 50% and over versus 25% or less in the other experiment.)

Another experiment by Kurtz, Walter, and Brenner (IFRR, 1950) compared a larger versus smaller number of added questions interspersed between sections of factual content presented by a film, but results for different groups were inconsistant and the results generally inconclusive. The question of optimum frequency of overt response, as well as the question of the relative importance under various circumstances of the role of response as (a) a motivational factor, and (b) a factor involving direct practice of a to-be-learned response (see p. ) is of considerable interest in relation to current alternative philosophies of programming autoinstructional material.

The approaches to programming associated with Crowder (1960) on the one hand, and with Skinner (1958, 1959, etc.) and his many followers, on the other hand, differ not only in the use of program branching and form of response required, but also in the frequency with which overt responses are required--or amount of explicit response as compared with the implicit responses involved in reading. As Lumsdaine (1961; 1962, HRCT, Ch. 12, p.615) has pointed out:

> This difference in response frequency is associated with the difference in the emphasis placed on the two roles of response. In Crowder's programs, the role of response is primarily one of testing to check on the knowledge which the student is supposed to have acquired through implicit responding in the course of reading a lengthy frame. Skinner and his followers assume, on the other hand, that the elicitation of appropriate overt responses by the student, with explicit reinforcement or confirmation of these responses thereby made possible, are *per se* a means by which the responses will be learned more effectively.

### Direct-practice effects versus side effects of active-response procedures

One of the objectives of a 1951 experiment by Michael and Maccoby (SRPI, Ch. 18) was to assess the direct effects of practice versus possible motivational effects that would show up in better attention and hence better learning on material not specifically practiced. Indirect evidence from the original World War II experiment by Hovland, Lumsdaine, and Sheffield (HLS, Ch. 9) suggested that both kinds of effects would operate, since the superiority of active over passive review groups was less when added external motivation was provided (by announcing that a test would be given after the lesson) than when the test was not announced. Michael and Maccoby gave practice through use of active-response questions on only half of the material covered in the film. They found marked gains on the material practiced, but no appreciable gains on the other half of the material. A similar finding was obtained in a later study by Kanner and Sulzer (SRPI, Ch. 28) using the phonetic-alphabet material as subject matter. These results point to practice as the key factor. The effect is presumably not to be limited to wholly "rote" learning, however, since in a later experiment, Maccoby, Michael, and Levine (SRPI, Ch. 19, pp. 295-309) found that the practice effects held even though test question wording differed considerably from that of the practice questions. In their later replication of this experiment (SRPI, Ch. 19, pp. 311-321), they used less interesting subject matter and found gains both on practiced and non-practiced material for groups who did not have the extraneous incentive of being told they would be tested. However, the latter gains attributed to motivational or incentive effects appeared to be somewhat smaller than the direct practice effects.*

One of the motivational functions that may be served by overt practice, especially when accompanied by prompt, reinforcing feedback, is that of keeping a student working at the learning task when time spent is on a voluntary basis. As has been noted elsewhere (Lumsdaine, HRT, Ch. 12, p.646):

This effect obviously cannot be observed in a fixed study-time experiment. Well-controlled experiments on this question are badly needed, even though positive results (more learning through more time spent in study) would raise the complicating question of where the time taken for the added study comes from. If it came from time otherwise wasted, the positive effect would be an important and clearly desirable one. If the added time used were "stolen" from other studies or other essential activities, however, a problem involving difficult value judgments would be brought into focus.

## Some Factors Influencing the Effectiveness of Active-Response Procedures

### Form of overt response in verbal learning

Skinner (1958, TMPL, pp. 137-158) has emphasized his conviction that the learner should compose his answer rather than merely recognizing it from a set of alternatives. Lumsdaine (1962, HROT, Ch. 12, p. 616) has noted that:

> This position is based on the assumption that there is low transfer from selective to constructive responding and that the presence of plausible distractors in a multiple-choice question will strengthen unwanted responses for some learners. This issue has important consequences for the design of teaching machines because of the much greater difficulty and expense in instrumenting a constructive-answer device that provides automatic answer discrimination. Skinner and others believe the latter to be desirable, even though they have compromised thus far, primarily through using the learner as the comparator in devices for teaching complex verbal subject matter. It should be noted that a straight comparison between existing free-response devices using student-compared answers versus multiple-choice devices with automatically compared answers would confound mode of response with type of feedback.

Gilbert (1959, TMPL, p. 476) semi-seriously predicted that experiments comparing multiple-choice and constructed responding were, or soon would be, "the most popular experiment current in psychology," and went on to indicate his reasons for questioning the value of such experiments. His prediction turned out to be in error, though he had an important point to make in offering it--see TMPL, pp. 476-477 and comments in Chapter 12 concerning contingent versus absolute hypotheses. [The most popular experiment (to which Gilbert's objection would also apply) has turned out to be comparison of overt and covert responding: See pp. 35-37, below.]

One of the experiments which has compared constructed and multiple-choice responding is reported by Coulson and Silberman (1959, TMPL, pp. 452-463), who also introduced--see below--experimental variation in the number of steps included in a program, and in the use of branching to vary program content. Constructed-response frames in the teaching program (which was a portion of a preliminary form of the program on behavior by Holland and Skinner, 1961) required the subject to fill in blanks in the sentences presented by each frame; for multiple-choice responding, the subject chose one from the two to five alternatives supplied for each blank. With either response mode, the correct answer to each item was revealed to the subject as soon as he had responded. Multiple-choice responding took significantly less time (44 minutes on the average) than constructed response (54 minutes). No significant overall differences were obtained using either multiple-choice or constructed-response criterion tests. However, with the criterion test requiring constructed responses, the constructed-response programs were superior under the non-branching condition. (No analysis was presented to assess this difference as related to the increased time required for constructed responding.)

In a second experiment on response mode, reported by Fry (1960, TMPL, pp. 469-474), paired associates (16 English-to-Spanish words and phrases) were used rather than continuous-discourse materials. There were three conditions, differing in the way time was controlled. The results with constructed-answer criterion tests (see also Chapter 15) showed the following significant results for both immediate and delayed tests: (1) under the first condition, working to a criterion of two correct responses at their own rate, use of constructed responses in training produced better results than use of multiple-choice responses, but took considerably more time (14 minutes versus 8 minutes) to finish the program. These results have the same ambiguity with respect to time vs. achievement as criteria as those of Coulson and Silberman (see Chapter 15). However (2), when time was controlled by letting students work at their own rate but giving them equal working time (Condition 2), or by arbitrarily keeping time per frame constant (Condition 3), use of constructed answers in training in Fry's experiment also produced significantly better results than multiple-choice responding. Another finding bearing on the response-mode problem has been reported by Zeaman (1958), who, for multiple-choice responding in teaching retarded children, reported better results when only two choices were provided than when three or more were used.

Overt and covert responding

In ordinary reading, listening to a lecture or watching a film, responses are implicit rather than being explicitly evoked by specific questions or gaps which the student is expressly directed to fill in. Experiments dealing with stimulus factors calculated to influence unidentified implicit or perceptual responses are described below. Explicit responding on cue need not, however, be overt (e.g., written or oral) but can be performed covertly or mentally. We will use the term covert response to designate responses which, unlike the implicit responding to a text or lecture, are deliberately made as explicit answers to a question or other express invitation for response, but which are not performed overtly. Such responses may serve as active symbolic practice, but afford a less clear basis for differential feedback from an instructional program or teacher than can be occasioned by overt responses.

One of the variables studied by Michael and Maccoby (SRPI, Ch. 18) was overt versus covert responding to the same set of questions. In their experiment, as in Fry's third condition, time allowed for these two kinds of responses was kept constant. They obtained no reliable difference between the effects of the two kinds of responses, even though the experimental conditions were sufficiently sensitive to detect fairly small differential effects. This finding was als replicated by Kanner and Sulzer (SRPI, Ch. 28) with paired-associate materials. A small-scale experiment by Evans, Glaser, and Homme (TMPL, pp. 446-451) compared overt and covert responding to a self-instructional program and failed to find significant difference in later test scores (see also Chapter 15).*

An interesting experiment on covert response was performed by McGuire (*SRPI*, *Ch. 27*). As in the above-cited Jaspen experiment, two rates of presentation were used, though rate was better defined in McGuire's experiment, which compared overt responses with explicitly occasioned covert ones (rather than overt participation versus no explicit response as in Jaspen's study). McGuire's subjects were given six practice trials in naming mechanical parts; time for these *practice* trials was constant for all subjects. The parts were also displayed and named in six *presentation* trials, which lasted two seconds for each part per trial in the fast condition, and four seconds each in the slow condition. Overt-responding subjects wrote the names of the parts, covert-responding subjects merely said the names to themselves. Despite the lack of any significant difference for the two forms of responding at the slower rate, *covert* responding was significantly better at the faster rate. One interpretation of these results is that if instruction has not adequately prepared the learner to respond correctly, forced overt responding may lead to distracting anxiety, to the practicing of errors, or both.

Another recent experiment concerned with response mode in the context of research on auto-instructional devices was conducted by Goldbeck (1960) to determine the effects on learning of three different modes of response in combination with variation in the level of difficulty of learning the responses. This study ~~which~~ is discussed in Chapter 15 in connection with the problem of the dual criteria of time and learning or achievement.

Results of recent experiments described by J. G. Holland (1960) and by C. McCullough and L. Van Atta (1960) in a symposium at the 1960 Convention of the American Psychological Association have also indicated that with carefully programmed pp. 134-151 continuous-discourse materials the performance of overt-responding students did not exceed that of students who read the same material with the blanks filled in. (In Holland's study, performance superior to straight reading was obtained only when the response blanks were chosen to represent the responses considered most relevant and were also well cued so that the student could readily come up with the correct response.) In commenting on these findings the symposium discussant, F. S. Keller, took the position that the data must be explained by assuming the occurrence of covert "textual behavior" and "built-in textual reinforcement."

As has been noted elsewhere (Lumsdaine, HROT, Ch. 12):

> Though such an explanation is necessary to account for the data in S-R learning-theory terms, it is obvious that such an explanation will be unsatisfying to many until a better theory of the functioning of implicit and other forms of covert responding is formulated in a way that can be subjected to experimental test. The experimental results to date certainly indicate the need for further analysis of the conditions under which overt and explicit responding result in optimal learning.

## Feedback, reinforcement and knowledge of results

Although there is a voluminous literature on the effects of reinforcement and variations in its scheduling on learning by infra-human species, relatively few experiments have studied reinforcement factors systematically as they operate in practical instruction of human Ss. Where this has been done, the role of reinforcement in the sense of a variable presumed to strengthen directly the effects of immediately preceding practice (in a Skinnerian, Hullian or Thorndikean sense) has seldom been disentangled from other functions of response feedback to the learner. (See Lumsdaine, 1962; Coulson, 1962 pp. 134-151). Among these other functions are indirect or cognitively mediated consequences of "knowledge of results," and the different ways in which knowledge of results may function in different kinds of learning tasks.

In the experiment by Michael and Maccoby (SRPI, Ch. 18), half of the active-response groups were, after responding to each question, told the correct answer; this was termed "feedback" or "KCR" (knowledge of the correct response). The other half of the actively responding groups were not given KCR. Instructional time was identical for the two sets of groups. The results showed slight though significant gains for the active-response procedure without KCR, but more marked gains (as expected) when KCR was provided. The results thus helped to isolate more clearly several components involved in earlier experiments in which similar feedback was invariably confounded with active response.

However, the precise functions of the KCR procedure still were not fully clarified since, as the authors point out, the procedure not only had potentially reinforcing properties (confirmation of initially correct responses) but also provided an opportunity for correction of errors and, in effect, for an additional implicit-practice trial in which covert practice of the correct response doubtless occurred. This ambiguity is also present in current use of a correct-answer panel, revealed after the student's response to each frame, in teaching-machine programs where the student serves as his own comparator. For further comments on this important point the reader is referred to the discussions by Lumsdaine (1962, Coulson, pp. 134- ); see also 1962, HRJT Ch. 12, pp. 618-620 .* In another study, Hirsch (1952, IFRR # 35) concluded that there was less effect when students were simply told right or wrong than when the correct answer was also presented (especially when the covert answer was presented in the context of repeating the question).

However, his results were far from clear-cut, due to the small size of the gains and large amount of variability.

It may be noted that the kind of "feedback" involved here does not seem to be quite the same as the differential feedback provided by the "knowledge of results" in a motor skill. For example, in dart throwing (cf. Judd, 1905) knowledge of results given varies with the accuracy of the antecedent response, and is used to redirect the subsequent overt practice; and the immediate perceptual response to this information does not in itself appear to provide a close surrogate for overt practice. However, in the kind of KCR provided in a Skinner teaching machine or through such procedures as those employed by Michael and Maccoby, the physical stimulus or information provided in giving KCR is the same regardless of whether the student responded correctly or incorrectly, though its interpretation differs in the sense of confirmation versus correction. The ambiguous role of KCR-feedback also enters into the theoretical interpretation of experiments on prompting and confirmation, to be discussed presently.

Little experimental research seems yet to have made a clear analysis of the factors in different kinds of response consequences; the general tendency seems to be to lump them all under the heading of "Knowledge of Results" or some other blanket term, thus confounding several factors. One comparison of the effects of different kinds of feedback has been reported in a study by Irion and Briggs (1957; see Briggs, *SRPI*, *Ch. 23*). They used a 20-alternative multiple-choice teaching machine (cf. Briggs, 1953; *TMPL, pp.* 299-304) for both paired associate and serial learning, and employed a fixed practice period of twenty minutes with several repetitions of the material, depending on each student's rate of working. In both tasks, showing the correct answer after each attempt was signficiantly better than a mere right-wrong indication. [ See also the earlier study reported by Stephens in 1953; (*TMPL*, *pp. 89-98*) , and the experiment by G. W. Angell (1949); see *HROT*, p. 621.] Some related experiments by Pressey (1950), Angell and Troyer (1948), Peterson (1931) and Little (1934) are reprinted in *TMPL*. Porter (1957; *TMPL* pp. 114 - 132 ) has written a critical commentary on some of these previous experiments, and has identified requirements for a more crucial test of the value of immediate reinforcement than they provide.

### Guidance or cueing of overt student responses

A principle underlying the design of most teaching-machine programs is that of building up mastery through small, easy steps, so that the learner is kept performing correctly from the very outset of training. (Cf. Skinner, 1958.) This means giving the learner a great deal of help initially through prompting cues, which are later gradually reduced or "vanished" so that the learner can get along without them. To date there is more evidence on the value of prompting as such than on the value of vanishing.

An early experiment on the use of prompts was reported by Pan (1926), who found that semantic hints indirectly suggesting the response term facilitated paired-associates learning of arbitrarily paired terms. One of the clearest pieces of evidence on the value of partial prompting or response guidance comes from a study by Kimble and Wulff (1953; *SRPI*, *Ch. 15*), who used a combination of film strips and workbooks to teach the reading of slide-rule scales, with a practice schedule that alternated short practice exercises with short segments of demonstration. The main experimental variable was the use of various forms of prompts. A prompted group was provided with a constraint cue which limited the possible responses, though without specifying exactly the correct answer, while a no-guidance group was allowed to make more errors, by withholding such prompts. There was a clear margin of superiority for the subjects in the prompted group, not only on test items used in practice but also in transferring to other similar items. The superiority seemed attributable to the choice of prompts that would help the student avoid types of errors shown by preliminary work to be the ones most commonly made.

Wulff and Emeson [SRPI, Ch. 30] report several small experiment. using eight relatively difficult paired-associate items, comprising four pairs in which the members of each pair were highly confusable. In one of the experiments, two paired-associate procedures were compared which differed only in that for one experimental group the stimulus materials were arranged in a fixed order to provide associative support through position cues used during early stages of training. The students using these materials performed better at later stages of practice in which the standard paired-associate arrangement was used (with scrambled order) than did the students who had used the scrambled-order arrangement throughout the entire period of practice. Several other interesting experiments are reported in the chapter in SRPI (see also the summary of these in HROT, Ch. 12).

The findings reported by Maccoby and Sheffield (1958; SRPI, Chs. 3-5) on optimal distribution of demonstration and practice are closely related conceptually to the question of prompting. The fact that practice immediately following short segments of demonstration was better than deferring practice until a larger section of demonstration had been completed (see below) is interpretable as favoring strong (temporally proximal) over weaker (temporally more remote) prompting in the case of procedural skills.

"Prompting versus confirmation"

Experiments reported by Cook and Kendler (1956) and also by Cook (SRPI, Ch. 21) used letters as stimulus terms and nonsense drawings as the required responses in comparing training procedures they termed "prompting" and "confirmation." On prompting trials, Ss were shown the correct response pattern after exposing the stimulus letter but before they were asked to respond to it. A comparison condition replaced the prompting trials with what the authors called "confirmation" trials. In these the subject was told to respond before being prompted by exposure of the correct response figure; then, after he had been allowed time to respond, the response pattern was exposed. The "confirmation" procedure thus provides either verification or correction of the response that the student has given. With both procedures, an unprompted and unconfirmed test trial was given by the experimenters after each three training trials. The condition using prompting trials was clearly superior to the confirmation procedu both in the original experiment and in later, varied replications (see SRPI, Ch. 21).

Some comments on these findings are given by Lumsdaine (1962, HROT, Ch. 12) p.623, who points out that:

> ...~~it is of interest to~~ we should note that the response used in the experiments was one which was fairly difficult to to learn (a nonsense drawing which had to be imitated

accurately) rather than one (such as learning English equivalents of foreign words) where the response term is already familiar but merely has to be hooked to the proper stimulus equivalent. Another significant aspect of the procedures used by these experimenters, in this and subsequent experiments, illustrates the difficulties of experimentation in which the process of measurement may alter the learning conditions which it is designed to assess: the use of interspersed test trials after every three prompted trials in these experiments meant that the prompting procedure was actually a mixture of prompted and unprompted trials (the latter used for measurement purposes, but obviously influencing the total learning program also).

Clarification of Cook's conclusions is provided in an experiment by Angell and Lumsdaine (1960; SRPI, Ch. 24), which used a partial degree of prompting as compared with complete prompting for using the substantially same learning materials and stimulus-response time intervals as those of Cook. (This experiment can be viewed as an initial investigation of the efficacy of partial or incomplete prompting; see also further comments by Lumsdaine, SRPI, Ch. 31.) The basic comparison in the 1960 Angell and Lumsdaine experiment, was between two groups, one of which received the prompting on all trials in learning a set of paired-associates, and one of which practiced responding without prompting on every fourth trial. The results indicated that learning was more efficient under the experimental condition of incomplete prompting (prompting on three-fourths of the trials) than under conditions of complete prompting (prompting on every trial). This result is in accordance with the theoretical predictions outlined in the paper by Lumsdaine (1959, TMPL, pp. 534-539), and is in contradiction to the interpretation derived from previous studies, that maximum prompting represents the optimal condition.

Interaction of prompting and overt responses. A further experiment ← undertaken by Cook and Spitzer using the same materials as in the above-cited experiments (see SRPI, Ch. 21 and TMPL, pp. 604-605), employed a 2 x 2 design in which the variable of prompting-versus-confirmation was covaried with that of overt practice-versus-no overt practice. Their results (in terms of the number of correct items on test trials) verified the predictions that the best condition would be prompting with no overt practice and the poorest would be overt practice with confirmation. From their data, Cook and Spitzer concluded that overt practice interfered both with learning the response term itself and with connecting it with its proper stimulus, while the delay in presenting the response term (in the "confirmation" condition) interferred with connecting the response to its proper stimulus, but had no consistent effect on learning the response term as such.

Results favoring straight prompting over confirmation-correction were found in the experiment previously reported by Briggs (SRPI, Ch. 23), who found that several methods of feedback following the student's response on each trial were inferior to a condition in which direct prompting prior to response were given on all trials. A complicating factor in the support given by these experiments to the use of maximum prompting is that the time intervals used did not preclude anticipatory responding before the prompting cue was given. However, Kopstein and Roshal (1955; SRPI, Ch. 20) found that in learning paired associates (giving Russian names for common objects) simultaneous presentation of S and R terms was clearly better than "staggered" presentation, in which the response term joined the stimulus term after an interval of about two seconds. (Overt responses to these presentations were not required; acquisition was measured by interspersed test trials following each third presentation of the paired terms.) This of course tends to support the use of immediate cueing to head off incorrect anticipatory responses. (The authors' speculation that the advantage of immediate prompting may drop out or even reverse in later stages of acquisition could not be checked from their data.)

*Transition or "vanishing".*

¶ In the Briggs experiment (*SRPI*, *Ch. 23*), a prompting condition was followed, for different groups, by each of several unprompted conditions. Lack of significant differences was attributed to the fact that a quite high level had already been attained by the end of the initial prompted period of practice. In any case, no condition was included in which the prompted mode of practice was used throughout. Evidence on the theoretically very important question of transition or "vanishing" in verbal learning is thus not provided from this study; the evidence to date comes mainly from experiments on filmic demonstrations for procedural training, reported by Maccoby, Sheffield, and associates (*SRPI*, *Chs. 3-5*). Although they used mechanical-assembly as the learning task, the principles they investigated are considered common to any case of complex sequential learning where use is made of demonstration coupled with practice. The theoretical factors involved include the cueing prompting of correct responses during practice (see Sheffield, *SRPI*, *Ch. 2*) and the integration or consolidation of elements into total task performance.

*Optimal lengths of demonstration and practice segments.* Two of these studies (*SRPI*, *Chs. 3 and 4*) were designed to test concepts relating to the optimal lengths of demonstration and practice for learning sequential tasks. Filmed demonstrations of each assembly task were divided into several segments, each of which was a "small step" which approximately 75% of the subjects could correctly perform immediately on seeing it demonstrated. Results, expressed in terms of mean "performance rate" (number of correct assembly responses per unit of time), showed that the poorest performance, both during practice and on later tests, was obtained when the entire task was demonstrated before it was practiced. In most of the experiments, the best performance was followed at once by practice of just that segment, before the next segment was demonstrated. However, results also indicated that it was even better to start the students with small demonstration steps, and then transition to larger steps in later trials. For superior students, the best results were obtained with a self-pacing procedure in which the student himself regulated the length of the demonstration segment. Under these conditions, the superior students tended to adopt the transitional pattern spontaneously, starting with short easy steps on the first trial but ending with the whole practice by the last trial.

*Repetitive versus continuous demonstration and practice.* Other studies conducted by Sheffield, Maccoby and associates (*SRPI*, *Ch. 6*) compared a "repetitive practice" procedure in which each of several sub-assemblies were demonstrated and practiced *twice* before moving on to the next sub-assembly, Versus a "continuous practice" treatment in which a sub-assembly was practiced once before demonstration of the next segment but with the entire demonstration and practice sequence then repeated. —

For one of the two mechanical-assembly tasks that were used, the results substantiated the hypothesis that the learning of a lengthy serial task can be facilitated by breaking it into "natural" (contextually similar) units, with adequate separate practice on the individual units being emphasized rather than practice of the task as a whole; but with the second task no significant difference between the treatments was noted. In interpreting this result, Sheffield and Maccoby (cf. SRPI, Chs. 6 and 9) point out that the potential advantages of the "repetitive" procedure stem from its tendency to reduce interserial interference. With an intrinsically well organized task like the second one they used, there may have been too little room for improvement in acquisition by procedures designed to minimize intraserial interference between elements of the separate subtasks. The advantage of the repetitive procedure would thus be expected to vary with the intrinsic organizational characteristics of the task being demonstrated.

## ← Other experiments on "size-of-step"

In the first of the above-cited Maccoby-Sheffield studies, the longer practice segments provide more remote, thus less direct cueing for the steps of the procedure; and the steps preceding practice are thus smaller both in length and difficulty. Amount of instructional material was a constant, whereas in verbal-learning studies the "difficulty of steps" has been varied by decreasing the number of steps and thus also the amount of instructional material. In a study using verbal learning, Evans, Glaser, and Homme (TMPL, pp. 446-451) varied the number of steps required to get through programmed lessons on number theory and music fundamentals. Data for a small number of cases indicated with small steps less time per step (though more total time) was required, and fewer errors occurred in the course of learning. Another comparison of long, numerous step programs (thus with easier, or "shorter" steps) in verbal subject matter has been reported by Coulson and Silberman (TMPL, pp. 452-468), who also varied "step size" by deleting certain steps from the program to yield fewer steps but with larger gaps separating them. Smaller and more numerous steps gave better learning scores, but these results, like those of Evans, Glaser, and Homme, are difficult to interpret because (a) the small-step program involved more practice on varied examples and (b) it took considerably more time.

For a discussion and interpretation of most of the experimental data on student response, (including that summarized above and a few of the studies to be reviewed in the next chapter) the reader is referred to the chapter by Lumsdaine (SRPI, Ch. 31).

# Chapter 14

# Some Experimental Studies of Stimulus and Organization Variables in Instructional Media

## Introduction

In this chapter we continue our review of instructional media experiments. The first section of the chapter deals with cueing or stimulus control factors aimed rather directly at controlling specific though generally implicit student responses, of a symbolic and perceptual kind, in relation to specific points in an instructional sequence. The second section deals with some characteristics of over-all arrangement or organization of instructional programs, aside from the kinds of relationships of practice to presentation already discussed in Chapter 13. A brief final section surveys some miscellaneous studies on verbal factors and some studies dealing with manipulation of motive-incentive-interest factors in instruction. The special conventions for reference citation introduced in Chapter 13 (see page 13-  ) are continued in this chapter.

## Stimulus Factors in Instructional Programs

### Stimulus control of implicit responses

A stimulus-response conception of the learning process (see Chapter 12) implies that, aside from explicit, instructed responses, learning is mediated through implicit or unidentified covert responses which take place while reading, listening, or watching a demonstration. It can be presumed that such responses are governed in one way or another by stimulus events which media provide, and which govern momentary perceptual or cognitive responses that mediate the development of skills, knowledges, and other desired outcomes of instruction. Thus, as implied in Chapter 2, many if not all independent program variables, in addition to those affecting explicit or specifically-occasioned student respones, can be viewed as providing suitable, effective control of implicit responses that mediate learning. This orientation may be useful in interpreting the further variables discussed in this chapter, including varidbles such as repetition, content organization, and pictorial-quality factors ~~that were~~ identified solely in terms of stimulus factors, as well as variables identified in terms such as attention-directing, familiarization, or discrimination.

Question-posing techniques. May and Lumsdaine (1958, Ch. 6) showed that the teaching effectiveness of existing film materials could be improved somewhat by splicing in "thought proveking" or "motivating" questions designed to arouse curiosity without calling for explicit responses. However, neither they nor Michael, Maccoby and Levine (SRPI, Ch. 19) found that questions preceding rather than following segments of the film to which they referred were as effective as overt-practice questions that followed the expository sections of the film. In a further study (M&L, Ch. 7) a film with purely expository narration was compared with one which made liberal use of rhetorical questions (which Ss were not asked to answer). Data for fifth grade pupils showed little if any improvement from using the rhetorical questions. While these results do not rule out the possibility of profitable use of rhetorical questions in a film script, the findings do suggest that introducing questions may be more effective if the pupil can correctly give an immediate overt or implicit response.

Perceptual factors

Sheffield (SEPI, Ch. 2) describes the formation of "perceptual blueprints" as a means of providing cues for the learning of lengthy sequential tasks. This concept was illustrated in an experimental technique called the "implosion" method, used in an experiment (SEPI, Ch. 8) using filmed demonstrations of assembly tasks. A rapid succession of "stills" of each part added into its appropriate place in an assembly made the parts appear to "jump into place" in sequence. Such an "implosion" device, which required only a very short amount of additional time, made a significant addition to the teaching of a lengthy mechanical task.

The advantage of depicting a to-be-imitated performance so that the learner observes the task from the same aspect which he will face when he comes to perform the task as compared with the demonstrator's facing the learner with the learner's view of task performance reversed left-to-right was demonstrated in an experiment by Roshal (1949, IFRR#5; see SEPI, Ch. 11) with a film on knot tying. In this experiment a "zero-degree" or "subjective" camera angle (showing the demonstration as the student would view the knot in tying it) was compared with a "180-degree" camera angle (showing the demonstration as the student would see it if he were watching a demonstration by an instructor who was facing the student). The reasons for the superiority of the "zero-degree" angle have not been fully explored, but it may be noted that the "zero-degree" angle presents to the learner more exactly the same cues as those to which he will have to respond when he is actually tying the knot.

The adequacy of implicit perceptual responses for guiding later practice could obviously be affected by the adequacy of the original perception. In demonstrating a motor-coordination skill where the skilled act may occur too swiftly to provide a stable perceptual pattern, the use of slow motion photography would seem likely to be especially useful. However, little experimentation has been completed on this stimulus variable, ~~although it~~ which represents a unique capability of motion pictures. The limited evidence to date comes mainly from a study by McGuire (SRPI, Ch. 13) who found some advantages for slow motion over normal-speed depiction in experimental films used as demonstration aids in a perceptual motor task.

# Silverman (1958) has compared "animated" visual aids with static ones as aids to verbal instruction given by a tape recorded lecture. With verbal tests, no significant differences between the effectiveness of using animated and static transparencies were found, but significant differences were obtained with performance tests. Roshal (IFRR# 5 ; SRPI, Ch. 11) found that a sound motion-picture demonstration on knot tying was significantly better than a parallel presentation that substituted a series of still pictures for the

motion picture. The advantage to be expected for motion depiction would, however, probably depend on the nature of task and on the frequency of transition and rapidity of change of the successive stills. Further experiments on this question are needed, particularly since they would have implications for equipment requirements that differ greatly in cost.

Considering the economic differences between using motion pictures and successive stills for instruction, it is surprising that little research has been done on the representation of motion in still pictures. One of the few studies is that by Smith and Resnick (1953), who showed that untrained observers could differentiate fairly reliably between live drawings of human figures that were excerpted from a motion sequence and those where the figure was not moving. No systematic experiments appear to have been done on the effectiveness of employing "swish lines," multiple exposures and other devices calculated to delineate motion in a still picture, nor on speed or frequency of cutting from one still to the next in sequence, as these affect perception or imitability of demonstrated operations.

In general, little experimental study has as yet been made of the specific characteristics of animation techniques. In a study of animation techniques by Lumsdaine, Sulzer, and Kopstein (SRPI, Ch. 17) student attention was directed to relevant cues by animated visual indicators accompanying narration in an instructional film on micrometer reading. Experimental films using these overlay-animation techniques were found, as expected, to be superior to control films which were identical except that the overlay animation devices in the key sequences in the film were omitted. Some methodological aspects of this study are further discussed in the next chapter.

An additional comparison was also made of the effects of two different presentations on the teaching of nomenclature. A treatment which employed pop-on arrows to indicate the parts of the micrometer named orally by the narrator was found to be less effective than a more complex animation treatment in which printed labels were also popped on, and the various portions of the micrometer were differentially stressed by fade-in techniques.

*Pictorial and auditory "embellishments."* The producers of the original version of the phonetic-alphabet slide film used in two of the above reported studies of overt-response procedures (*HLS*, *Ch. 9*; *M&L*, *Ch. 5*) evidently believed that the musical accompaniment and the humorous cartoons and comments employed in the film would make it a better teaching device than unadorned presentation of the material. This could have been accomplished if (1) there were a need for a general interest-getting function which these features might have served; (2) if these features had served to direct attention to relevant cues related to the learning task; or (3) if such features had been designed to serve as memory aids (mnemonics) so as to facilitate association of cue (letter) and required response (phonetic name). However, the results obtained by Lumsdaine and Gladstone (*M&L*, *Ch. 5*) showed that the "fancy" presentation using the pictorial and auditory embellishments was not only not more effective but was actually less effective than a "plain" version in which the letters and equivalent phonetic names were merely presented starkly on a plain background with neither pictorial nor auditory "fancying up."

These results suggest either that none of the three hypothesized functions was operative or else that they occurred to such a small extent that, for the training conditions found in the experiment, they were overbalanced by other, detrimental effects of the comic embellishments. The latter could include drawing attention away from rather than toward the essentials of the learning task, by distracting the learner from implicitly rehearsing to himself the association between cue (letter) and response (phonetic word) that needed to be established to master the task as defined. Under certain conditions, however, the possible positive effects of the embellishments might more than offset such a distracting influence. This could easily happen, especially if attention to what is being presented (insuring "exposure" to the learning materials) were the most critical requirement; the presence of the entertaining material might well make the difference between exposure (watching) and nonexposure (e.g. turning the television set to another channel or going to sleep), thus fostering at least some learning, rather than no learning at all. (For further discussion, see Lumsdaine, HROI, Ch. 12; or Lumsdaine, Sulzer, and Kopstein, SRPI, Ch. 17.)

Experiments on pictorial "quality." May and Lumsdaine (M&L, Ch. 2) compared a crude, preliminary pencil-sketch version versus a finished, regular, color version of an elementary science film on the causes of seasonal change. Although this was a well-controlled experiment as a basis for determining the relative amount learned from these two specific presentations, the content of the two versions was not entirely parallel. The results failed to show an appreciable difference between the effectiveness of the two. However, because of the imperfect control of content, these results can be interpreted, at most, only as showing that such a crude presentation as this one may at least equal in teaching effectiveness a polished color film costing more than ten times as much to produce. The results certainly do not demonstrate that technical quality is necessarily an unimportant factor in the effectiveness of an instructional film. However, in a later study reported by Zuckerman (1954, see summary in SRPI, p. 527 ), a story-board-based filmstrip of about 250 frames was used which closely paralleled a finished film. As in the May and Lumsdaine study, the visuals in the filmstrip were relatively crude sketches, as contrasted with the combination of live action and elaborate animation used in the final color film. The purpose of this study was to determine whether point-by-point effectiveness of the final film could be predicted from the filmstrip mock-up, but the results also replicated those of the May and Lumsdaine study in showing (1) wide variations in the effectiveness with which either presentation conveyed specific points of content and (2) no material over-all superiority for the polished color version over the crude visual presentation.

Color. No really definitive studies have been made on specific ways in which color may contribute to learning from instructional media. The principal experiments (M&L, Ch. 2; VanderMeer, 1952, IFRR#28) have compared black and white versus color prints of the same instructional films. (See Ch. 3, p 27f. and Ch. 6, pp.13-16.) These studies failed to show significant differences in learning favoring the color films, though some of VanderMeer's data indicated slightly superior retention in the case of some of the color films be used. These experiments lend little support to those who argue for producing classroom teaching films in color in order to achieve greater educational effectiveness. But neither do they justify the conclusion that color is wholly irrelavant to learning, even in films which have no intrinsic color content to be learned (see further discussion by Lumsdaine, M & L, Ch. 2, pp. 29-30).

### Pictures versus words as stimuli in paired-associate learning

A theoretically-oriented study of the functioning of pictures as compared with words was conducted by Lumsdaine (see M & L, Ch. 10) in teaching paired-associates. The experimental question was whether using pictures or words made for better learning, (a) in presenting the first or stimulus object in the piar, and (b) in presenting the second or response term. The results showed that the use of pictures to represent the first object of the pair was better than verbal presentation, but that verbal presentation (printed words) was definitely superior to pictorial presentation for the second ("response") term of the pair. (This was true both for college and grade school subjects, for individual as well as classroom presentation, for brighter learners as well as the slower ones, and regardless of a number of systematic variations in condition of presentation.) This finding was subsequently replicated by Kopstein and Roshal (SRPI, Ch. 20) who used Russian-English vocabulary terms and also found that learning under the picture-word condition tended to produce better performance with verbal as well as pictorial test stimuli.

Other visual-display factors. In a widely-cited study by Swanson (1954) several different devices were used as visual aids supplementing lectures to familiarize experienced mechanics with subsystems of an Air Force bomber. No significant differences were found in the effects attributable to use of mock-ups, tow dimensional symbolic diagrams, and realistic pictorial charts. A later experiment by Swanson, Lumsdaine, and Aukes (1956) used inexperienced airmen as subjects and two training aids (mock-up and diagram) from the earlier experiment. In this case, the mock-up group achieved superiority on a parts-recognition test; on a test of functional inter-connections the diagram group was superior, and on ~~the~~ a third test (verbal knowledge) there were no significant differences among any of the three groups. The conclusion was that the effectiveness of each training aid depends on specific instructional objectives so that these objectives would dictate which aid (if any) would be more effective.

A positive contribution for visual aids accompanying a recorded lecture was also shown by an experiment by Aukes and Simon (1957), who were concerned with the possible distracting role of irrelevant cues furnished by material extraneous to the point under consideration. They found that, as visual aids for a tape-recorded lecture, a standard equipment display was inferior to an "add-a-part" display in which only those portions relevant to each major point of the lecture ~~was~~ were present during discussion of that point.

"Live action" and realism

In an experiment reported by May and Lumsdaine (M & L, Ch. 3) a narrated version of a film on the causes of the seasons was compared with a story-dialog version for fifth and ninth grade students. The differences between the mean test scores of the two groups were too small to be reliable, although fifth graders did express significantly higher interest in the story-dialogue form of the film.

### Live versus filmed Instruction.

Shettel *et al* (1956) dealt with practical question of whether filmed recordings of lectures could serve as substitutes for the more expensive and cumbersome mobile training units then commonly used by the Air Force for the training of mechanics. In each of three replications, live-lecture groups employing various visual aids were compared with corresponding filmed-lecture groups. On the immediate (verbal) post-test, two of the live-lecture groups scored higher than the corresponding filmed-lecture groups, but the differences, though reliable, were small. On the delayed test, given six to eight weeks later, scores were much lower, and no reliable differences were found between any of the live and filmed-lecture groups. However, for one of the replications, the film was reshown, as a demonstration experiment, just prior to the delayed test. This predictably produced a reliable improvement in the scores on that test. This fact was used to emphasize the logistic advantage that, unlike the live lectures, the filmed ones could not only be kept available for periodic review, but could readily be scheduled immediately prior to the time their content was needed in operations--thus avoiding large decrements due to forgetting over the period between instruction and application.

## Visual and Auditory Channels

Cheatham (1950) has analyzed variables affecting the probable efficiency of visual and auditory modalities as communication channels. His conclusions and the survey of literature by Day and Beach (1950) make it apparent that the problem is a complex one that involves a number of variables. Where the visual channel is crowded, it is no surprise to find --e.g., the pilot study reported by Hoehn and Lumsdaine (1958) -- that audio task instructions can be much more efficient than printed ones. But in more complex situations such as class instruction on complex subject matter, it is unlikely that either modality would be consistently superior. The inconsistent results obtained by Nelson and Moll (1950) in comparing presentation by audio and visual elements of several existing films are thus not surprising. What is needed here are studies that test specific, hypothesized advantages for one or the other modality as a function of other instructional variables.

## Factors in the Content and Organization of Instruction

### Pre-instruction motivational and attention-directing procedures

Use of a pre-film test to identify points to be learned in a program has been studied by several investigators. In the animation study reported by Lumsdaine, Sulzar, and Kopstein (SRPI, Ch. 17) giving a brief pre-film test of five sample micrometer readings significantly increased the amount learned from the film. Several experiments reported by Lumsdaine (M & L Ch. 7) also showed that the amount learned from certain aspects of a film could be increased by procedures for directing the attention of students in advance to these aspects of the presentation--e.g., by oral or written instruction, or by giving a prefilm test. However, the increase in amount learned from the material covered by the pre-test tended to be offset by losses in the material not covered. In an unpublished 1953 study (see summary in SRPI, pp. 522-24) using a film on types of map projection, A. K. Romney, also reported that instructions designed to induce a "set" to learn facts significantly increased the scores on factual items. On the other hand, inducing a set to learn "principles" resulted in nonsignificantly higher scores on "principle" items.

A relevant pre-instruction procedure thus can have a selective attention-directing function, stressing certain aspects of content differentially. On the other hand, merely announcing that a test (of some sort) will be given following instruction can serve a general motivating function which in the 1949 phonetic alphabet study by Hovland, Lumsdaine and Sheffield (HLS, Ch. 9), had a significant effect in increasing learning. The procedure of announcing a test did not work, however, when Michael and Maccoby (1953; SRPI, Ch. 18) tried to use it for a similar purpose: making the announcement had no significant effect (probably because, unlike the Army subjects of Hovland, Lumsdaine and Sheffield, their students were already about maximally motivated). Effects of tests interpolated between two sections of film instruction for set-producing or motivational purposes were reported by (Kimble 1956, see summary in SRPI, pp. 514-17). Following preliminary film instruction on reading of slide-rule scales, the experimental group was given a test (without knowledge of results) prior to follow-up instruction both on scale reading and on the use of the slide-rule for multiplying and dividing. Results of the interpolated test were specific, as compared with control groups who had no interpolated test: the experimental groups showed greater gains on the kind of content covered by the test (scale reading) but no appreciable effect on the content not covered (arithmetical operations). The psychological basis for the effectiveness of pre-instruction or inter-session tests to direct attention to particular elements of a subject matter seems to be closely related to the more localised attention-focusing effects that can be produced by devices used within the program--e.g., the animated arrows and labels used in the above-mentioned study by Lumsdaine, Sulzer, and Kopstein (SRPI, Ch. 17) for directing attention to key aspects of the picture and for interpreting the significance of cues. Gains in effectiveness seem, in either case, to depend on selective focusing of attention on particular aspects of the material, so as to make those aspects stand out from competing possible perceptions. The successful use of techniques for doing this contrasts with the negative results of Neu; (1950, IFRR # 9 ), who failed to find any significant effects for "attention" devices that merely aroused attention in a general way, rather than directing it expressly to relevent cues or specific aspects of content.

## Familiarization

A number of studies have been conducted on the value of so-called familiarization training. Most of these merely show that additional time devoted to some form of familiarization produces beneficial results, without showing that using it is more effective than alternative ways in which the time could have been spent. For example, Weiss and Fine (1955; see summary in *SRPI, p. 525*) found that prior familiarization training on key terms in a film on the United Nations yielded scores on a post-film test which did not differ significantly from those achieved by the about equally time-consuming procedure of giving two showings of the film (though they were reliably superior to those produced by a single showing). One exception, which also deals with specific cue factors involved in familiarization procedures, is a study by Wulff and Kraeling (*SRPI, Ch. 10*), who noted that generally it is not necessary for the learner to attend to *all* of the features of a task to be learned; rather that he needs to learn to attend only to certain features important in performing it. Accordingly, it was assumed that "familiarization" procedures focusing attention on the important features of each part in a mechanical assembly (automobile ignition distributor) would facilitate learning to assemble the parts properly; students not given such "familiarisation" training would be very likely to attend to irralevent features during a demonstration of the assembly, and would thus be handicapped when they subsequently attempted to assemble the mechanism. Two familiarization treatments were compared with each other (and also with training without familiarization). In one treatment, subjects were trained to attend to important features of all the parts before any assembly training (accomplished by means of a 16 mm. sound film); in the other familiarization procedure (which took less time) the same features were pointed out just at the time when the assembly of each part was to begin. Although both familiarization procedures were shown to be effective in reducing errors, the most interesting finding was that giving all familiarization prior to assembly training yielded poorer performance, with respect to both selection errors and assembly errors, than did the other familiarization procedure (that took less time), in which familiarization was given in context, just preceding the utilization of each task element. Such specifically oriented studies seem more likely to be rewarding than more general evaluative studies of less well defined "introductory" and/or "summary" sequences such as those compared in the early study by Hovland, Lumsdaine, and Sheffield (HLS, Ch. 5) and later studies which have compared the relative effectiveness of summaries and introductions to several films or asked the equally ungeneralisable question, "Do introductions help?"

### Repetition, redundancy, and review

Amount of repetition. It is clear that learning may occur in a single trial (cf. Guthrie, 1935 as well as recent papers by Rock and others). However, the frequent need for repetition of material to assure mastery is well established; the concern of psychologists with "learning curves" or "aquisition curves" is in itself evidence of this general phenomenon. Experiments on repetition in instructional media have reflected a prevalent educational philosophy which tends to grade students on how much they learn in a given time rather than addressing itself, as has been not uncommon in laboratory psychology, to the length of training (e.g., "number of trials") necessary for instance to achieve some predetermined criterion of mastery.

Experimental data have in general reflected varying degrees of incompleteness of learning resulting from various degrees and forms of repetition. Instructional films seem generally to be designed with insufficient repetition (and/or too fast a "pace" or "rate of development") for most students to approach mastery on specific knowledge or skills if used without supplementation. For example, Jaspen (1950, IFRR, # 11) found that a "succinct" treatment, chosen as representing the pace and level of repetition commonly employed in many films was far from adequate to teach an assembly task, reasonably successful teaching of which required a much longer film. A common finding is that students achieve far from perfect scores on criterion tests after seeing an instructional film. This has led several experimenters to study the effects of repeat showings of a film in toto, and a few experiments have been conducted on varied amounts or arrangements of repetition within an instructional film or other program.

Typical findings, predictably, are that two showings teach more than does just one (e.g., Kurtz, Walter and Brenner, 1950 IFRR, # 16). Additional showings may or may not show significant further gains. McTavish (1949; IFRR, # 12) reported that "...the contribution made by repetition of showings fell off rapidly after the first repetition...only the increment attributable to the first repetition is statistically significant." (p. 7) McTavish's study and the later study by Lumsdaine, Sulzer, and Kopstein (SRPI, Ch. 17) thus indicate the predictable existence of a "saturation point," although one study dealt with repeated demonstration within a film and the former with repeated showings of films. However, the important thing is not the mere existence of a "saturation point" but the determinants where this point will occur. It should be obvious that there will be no fixed, low value for the number of repetitions that can be usefully employed. This is illustrated in comparing McTavish's works with those of a study reported by Cook (SRPI, Ch. 21), using films on map symbols with one, two, and three repetitions of a key review portion of the film. Results showed benefits with up to three repetitions though with amount of improvement decreasing with additional repetitions. A. K. Romney (see SRPI, pp. 522-524) also found evidence to indicate that, under certain conditions of learner "set," three repetitions of a film were superior to two. On the other hand, McGuire (SRPI, Ch. 13), found no significant difference between the effectiveness of a long film which contained four repetitions of each instructional point in a demonstration on pursuit rotor tracking and that of a shorter film which presented only two repetitions of each point.

Lumsdaine, Sulzer, and Kopstein (SRPI, Ch. 17) studied the effect of increasing the number of examples in instruction on micrometer reading. From three (the number included in a standard "training film") to ten examples were employed. The potency of repetition was evidenced by the fact that adding only three additional examples produced roughly twice as much gain as did the use of overlay animation techniques employed in the original film. Six or seven examples were much better than three; ten examples were somewhat better than six or seven. Gains from using added examples were understandably greater for the harder-to-read kinds of settings. The more intelligent students benefitted more from added examples than did less intelligent students.

Types of repetition. The question of the relative merits of varied as against identical repetition was investigated in a study by Kanner and McClure (1956; see summary in *SRPI, pp. 510-512*), who used films on micrometer reading to compare the use of four examples, each seen twice, versus the use of eight different examples, each seen once. In a first experiment they failed to find a statistically reliable difference between the two treatments immediately after training, but did find a reliable, though small, difference in favor of the varied-repetition treatment for the more able subjects, when tested a week after training. In a second experiment, however, no statistically reliable difference of any kind were found. On the other hand, results of an experiment by Rimland (1955, *IFRR, #45*) showed a consistent advantage for two identical repetitions over two different presentations (of about equal separate effectiveness) in teaching a perceptual-motor skill (knot-tying) by filmed demonstration. However, since a *specific* skill was being taught here, this finding is not necessarily incompatible with a hypothesis of greater effectiveness for varied examples in other kinds of tasks, such as micrometer reading, which require generalization from examples that are used to teach a conceptual or procedural pattern.

*Distribution of review.* Miller and Levine (1952; summary published in *SRPI, pp. 519-521*) compared the relative advantage of spacing review sequences through a film on Ohm's law so that a review immediately followed each of four major topics, versus "massing" all the review sequences at the end of the film. The comparisons showed the massed review to be clearly superior to the spaced review. A later replication failed to achieve significant differences (though under quite unfavorable circumstances: see *SRPI, p. 521*). A further replication (see *SRPI, pp. 522-524*) used a more interesting film 6n atomic attack) with high school subjects. For the spaced-review condition, the film was divided into four sections and each section was reviewed by means of recorded narrations and colored slides before proceeding to the next section. For the massed-review condition, the entire review was deferred until the end of the film. The massed-review condition was found to be superior to the spaced-review condition ($P = .02$, one tail). However, a still further varied replication (see SRPI, pp. 518-19) once more yielded no significant differences when an "active-participation" type of review was substituted for passive review. (See HROT, pp.645-6 and SRPI, pp.520-21 and 495-96 for possible explanations of these results.)

Organization of Instructional Sequences

Fatigue and distribution of effort. An experiment on distribution of effort (as distinguished from distribution of practice) was reported by Ash (1949, IFRR, #3 ). Using two film series, each consisting of four 15-minute silent films, Ash failed to find significant differences in effectiveness between showing all four films in a single, one-hour session and showing the films in two 30-minute sessions or in four 15 minute sessions. Effects were assessed in terms of retention one or two weeks later. On the other hand, Faison, Rose, and Podell (1955; see summary in SRPI, pp. 506-08), showed that the effectiveness of a single 24-minute film during a single showing could be increased by short (30-seconds) rest breaks to increase audience attentiveness. Use of infrared photography as an audience-observation technique indicated that the insertion of three such rest pauses in a 20-minute film led to reliable increases in an index of audience attentiveness and, in turn, to reliably greater learning as measured by a post-film test. There was a progressive, gradual decline in attentiveness for the "straight-through" showing, as contrasted with the abrupt rise in attentiveness following the rest pauses under the experimental condition.

Placement of practice in relation to task structure. A study by Margolius, Sheffield, and Maccoby (SRPI, Ch. 7) was conducted to test the hypothesis that, in learning a lengthy assembly task by demonstration and practice, the optimum placement of overt practice is at the completion of each "natural unit" of a total assembly task. A "natural unit" was defined as a portion of a task sequence which has its own distinctive context cues, such as a separate subassembly. Use of overt practice at the end of such "natural units" was compared with use of overt practice at the end of arbitrarily defined segments of the task (which, however, matched the "natural units" in length, though not in placement). Results for the film that was used showed no effects of the difference in treatment, but results might well differ for a task that was inherently less well organized than the one used in this study. (see SRPI, Ch. 7 and Ch. 9, esp. pp. 125-129).

Structuring by subtitles and outlines. In the previously mentioned study by Miller and Levine on spaced versus massed distribution of review (1952; SRPI, pp. 519-21), a second question was how the effectiveness of the film might be affected by frequent subtitles designed to identify, set off, and structure the various topics and subtopics (electrical current, voltage, resistance, and Ohm's law) taken up successively in the film. No significant differences were foundamong three treatments employing (1) major subtitles only, (2) "complete" subtitling for each subtopic, and (3) no subtitles with each sequence of material following the next without a break. That the lack of differences may have been due to the materials in the Ohm's law film being already well structured is suggested by the results of a later study by Northrop (1952; FRR #33). Northrop reported significant differences favoring film versions in which structuring outlines were employed as compared with similar films lacking these aids to organization.

Sequencing of instruction to facilitate discrimination. An illustrative study on the sequencing of paired-associate materials is the paper by Wulff and Stolurow (1957; SRPI, Ch. 29) dealing with "class descriptive" cues that differentiate subgroups of materials which a trainee must learn to identify. The task they used was the system for coding aircraft rivets by color markings, etc., in terms of four properties: length, diameter, head shape, and material. The application of the code was taught by presenting many different individual rivets, one at a time. Two forms of organization were used: (1) "Class organization" in which displays showed all the items in relation to one class of differentiating cues, e.g., rivet-head silhouettes or rivet-head markings that indicated diameter; and (2) "Item organization" in which displays gave the same information, but with all features about a given item presented at one time, e.g., a single rivet with each of the four coded characteristics depicted. With the latter method, it was theoretically unlikely that the learner could utilize class cues effectively during learning; thus it was predicted that class organization would be more effective. The results supported the theoretical analysis; that is, the method of presentation which fostered utilization of class cues proved reliably superior to the other method.

A second example of studies on sequencing of materials is the report by Detambel and Stolourow (1956). They were concerned with a problem-solving situation in which a varying display was presented during the course of successive trials. The trainee's task was to learn not only which of the components of the display was relevant to decisions he had to make, but also which rariation in the relevant component indicated that a particular decision was required. A training sequence with displays arranged to provide "minimum synchrony" was compared with a "maximum synchrony" display. In synchronous sequences, trial-to-trial changes occurred concurrently (synchronously) with respect to both relevant and irrelevant aspects of the display, so as to provide no way for the subject to isolate relevant from irrelevant features. In "asynchronous sequences", by contrast, stimulus components changed and irrelevant ones did not, or vice versa (irrelevant components changed and relevant ones did not); here the relevant changes thus were isolated from the irrelevant ones. Minimum synchrony between relevant and irrelevant components produced significantly larger proportions of solvers. though only at the lower ability levels.

The study by Rothkopf (1958) provides a third example of experimentation on sequencing of material for discriminative learning. He performed several experiments employing Morse code signals asstimulus terms, and their alphabetic or numeric equivalents as response terms. For example, one experiment compared practice orders involving the maximal separation of similar stimuli with orders involving maximal clustering of similar stimuli (see HROT, p. 648, concerning other experiments). Results indicated that maximal separation of similar stimuli in the practice order produced the more efficient learning. However, in an earlier study Gagné (1950) had found that successive presentation of similar stimuli was a more effective practice sequence in a paired-associate task which employed nonsense figures as stimulus terms and nonsense syllables as response terms. The reversal of the Gagné findings by Rothkopf's results is indicative of the complexity of factors governing optimum sequence in programing. Thus, optimal sequence is likely to vary as a function of such factors as susceptibility of stimulus materials to use of mediating descriptive responses by S, and the amount of response-term learning required.

## Other Factors in Instruction

Changes in effects with passage of time. It is commonly recognized that effects of instruction may change or dissipate with lapse of time. The most familiar phenomenon is that of decreased retention due to "forgetting". One of the practical advantages of reproducible media such as films and self-instructional programs is that their use can be phased so as to reduce the losses due to forgetting by scheduling them shortly before the time when the material learned is to be put to use. This advantage is brought out in the demonstration experiment by Shettel et al. (1956), in which most of the material learned from conventional instruction had been forgotten after an interval of several weeks—a loss which was prevented or offset by showing the filmed version of the instruction just before its use in testing or practical application was demanded. An important exception to the general expectation of diminishing effects with passage of time is the finding by Hovland, Lumsdaine and

Sheffield (1949, Ch. 7) that some of the effects of films on attitudes or opinions showed an increase with passage of time (in contrast to effects on factual knowledge, which showed the usual phenomena of forgetting). As factors in the sequencing and organization of instruction begin to be attacked in more thoroughgoing fashion by study of alternative programming techniqes, it is to be expected that other instances will be found in which the immediate superiority of a particular instructional-sequence treatment may not necessarily be predictive of its relative effectiveness in later or transfer testing. This might well be found, for example, with respect to distribution of review (see the previous discussion of the 1952 results of Miller and Levine).

Some verbal factors in instructural materials

"Readability". Considerable attention has been attracted by so-called readability formulae propounded by Flesch (1946, 1949), Dale and Chall (1948) and others. However, relatively little effort seems to have been devoted to experiments to test the utility of the rules implied by these formulae for simplification of language to increase the effectiveness of instructional media. For the most part, studies (e.g., Park, 1945; Chall and Dial, 1948) have compared existing prose passages found to differ in terms of criteria of difficulty, rather than manipulating material by application of the "rules" and then experimentally comparing the comprehension of a resulting simplified version with a more difficult version. One exception is the study by Gladstone (M&L, Ch. 4) who, rather than dealing with printed presentation, constructed alternative forms of the narration for a 20-minute instructional film by invoking rules for style simplification. The two commentaries had the same factual content, but differed in difficulty level in terms both of the Flesch formula and the Dale-Chall formula. The results of experimental comparison substantiated in general the expectation that more factual material would be learned from the easier commentary. A study of similar sort was that of Allen (1952), who co-varied interest and the difficulty levels of the commentary independently and found results generally similar to those of Gladstone. Much further experimentation will be needed to determine what aspects of the simplification procedures used are the most effective ones.

*Importance of verbal component in audio-visual media*. Studies that seek to investigate the contribution of audio accompaniment to instructional films may be difficult to interpret because the assignment of information to audio or visual modes is rather arbitrary in the first place. For example, the findings of Nelson and Moll (1950; *IFRR* #18) concerning the relative amounts of learning attributable to picture and sound track in certain films are inherently questionable in generality because they can be taken as reflecting how much of the information tested by the experimenters happened to be given in the picture and in the sound track in those particular films. In the demonstration of motor skills the appropriate content for visual material and accompanying auditory commentary is perhaps more unambiguously generated by the nature of the task itself. McGuire (*SRPI*, Ch. 13) using a demonstrational film designed to teach pursuit rotor skill, found that film plus descriptive narration produced reliably better results than the filmed demonstration alone. An experiment by

Thomson reported by May (1946; see M&L, Ch. 12) studied a different kind of verbalization, used as an aid to the learning of an assembly task. It was found that guided verbalization of steps by the learner facilitated performance if the verbalization was relevant, but hindered performance if the verbalization was irrelevant. The importance of *relevance* in using verbal labels is also shown in studies that have investigated the utility of teaching nomenclature. For example, Saltz and Newman (1960) reported an experiment to test the hypotheses that (a) if the symbols used in the instructions are to be meaningful, they must be connected to the appropriate acts and equipment components to which they are referents (component-name learning), and (b) if the symbols are to aid performance, they must be connected *to each other* in the order that corresponds to the assembly sequence (order-of-name learning.) The amount of component-name learning and the amount of order-of-name learning were manipulated independently. Results were as hypothesized but also indicated that with no order-of-name learning, small amounts of component-name learning facilitated assembly, whereas large amounts of component-name learning produced decrements in assembly performance. These and related findings have been summarized as follows (Lumsdaine, *HROT*, Ch. 12):

> We may conclude that one way in which nomenclature *can* be effectively used is to utilize its cue properties, but that special attention must be given to seeing that a way is provided to these cues can operate effectively. For example, in facilitating assembly, sequential ordering of the verbal cues appears necessary in order for the cues to function

effectively in guiding sequential performance. This can either be assured by an explicit training procedure, as in the Saltz and Newman experiment, or be selecting verbal cues that already possess an inherent order as in the previously mentioned study of Thomson [M&L, Ch. 12] in which numbering the parts in the order they had to be used in the assembly considerably facilitated performance.

Characteristics of narration. "Rate" of verbalization and other factors in style of verbal commentary accompanying demonstration films has been studied in several experiments. Jaspen (1950; IFRR #17) and Zuckerman (1949; IFRR #4) varied the rate of verbalization in commentary accompanying demonstration films and concluded that a "medium" (90-130 w.p.m.) rate of verbalization was preferable. However, rate was to some extent necessarily confounded with content and redundancy, as the commentary rate variation was in the context of a fixed-pace demonstration; also, it seems likely the specific rate that would be optimal could well vary considerably depending on the informational requirements of the task. Zuckerman also compared various modes of narration differing in voice and mood and concluded that imperative instructions were superior to passive-voice descriptions of the steps in the task being demonstrated.

Organization of arguments in persuasive messages. Mention is warranted here concerning a number of experiments on factors in the sequencing of verbal material in persuasive communications. An extensive group of studies in this field is that of Hovland and his co-workers (e.g., Hovland, Janis and Kelly, 1953; Hovland, 1957). These studies illustrate the evolution of findings and re-delineation of problems for research in a programmatic effort on the effects of verbal communication, even though they might be considered tangential to the design of specifically instructional materials. However, as behavior and communication theories evolve, some of the theoretical substructure for effective practice in persuasive communication seems likely to show points of close similarity with that underlying communication designed to influence knowledges and skills. For example, there is an interesting parallel between the rules used for sequencing arguments in the communications used to study one-sided versus two-sides arguments (HLS, Ch.8; also Lumsdaine and Janis, 1953) and the rationale used in anticipating and heading off anticipated "small step" self-instructional programs. Thus the investigator concerned with purely instructional media may increasingly need to become informed about the substantive findings and theoretical developments in research in persuasive communication, as well as examining such studies in terms of relevant methodological points they involve.

## Interest and motivation as independent and dependent variables.

In the design of instructional media, interest and motivation may figure either as independent or as dependent variables. In terms of independent variables we ask how manipulation of motivational aspects of an instructional program may increase the extent to which desired instructional objectives (of whatever sort) are attained. In terms of dependent variables, we ask how manipulation of instructional variables (of whatever sort) may not only affect knowledge and skills, but also influence motivations or interest. The latter outcomes may be desired for their indirect value in facilitating or stimulating further, subsequent learning, and this later behavior may thus be the real dependent variable.

An example of a study where interest was used as a dependent variable is that reported by Lumsdaine ( M & L, pp. 31-45), which compared two versions of a film presenting scientific explanation of the causes of seasonal change. One version of the film used live dialogue and the other used narration. Although the amount learned from the live dialogue version was not found to be significantly greater than that from the narration version, the fifth grade students, did tend to like the live dialogue version more than the narration version, and those that were shown the live dialogue version expressed somewhat greater interest in the subject matter of the film than those who were shown the narration version.

The relationship between interest in an instructional presentation and amount that is learned from it have not pinned the question down very closely by experiments to date. In one suhh experiment (HLS, Ch. 5) an "interesting" kind of presentation (dramatic) was compared with a more straightforward, presumable less interesting kind of presentation. This experiment did not provide really analytic definition of the specific variables involved, and it yielded no significant difference favoring one treatment over the other. VanderMeer (1953; IFRR #50) made a somewhat similar comparison on the effect of two versions of a film on personal-hygiene habits. A straight leature version and one which had been "jazzed up with folk music in elaborate Hollywood style" (Cook, 1960, p. 92) roughly equal changes in personal hygiene behavior. In both of these two studies, two films representing approaches that presumably differed with respect to interest were compared, but without precise manipulation of specific, identifiable factors and close control over other aspects of content.

Feedback of test results to students in the previously-mentioned experiment by Lumsdaine (M & L, Ch. 7) involved motivational as well as informational functions. An experiment in which reports of test results were confined to motivational aspects was reported by Kimble (1955; see SRPI, pp. 512-514).

The effects of praise and reproof were compared for students' performance on a test covering preliminary phases of an instructional topic. In terms of subsequent test performance on material learned before the administration of praise or reproof there was no significant effect. However, on material taught after the administration of priase or reproof, the reproved group scored highest, the praised group next, and a no-incentive-comment group lowest. The latter differences were significant; so that the results point up the questionable generality of the earlier widely-cited results of Hurlock (1925) which have generally been advanced as evidence favoring the use of praise rather than reproof as a motivating agent.

McGuire (SRPI, Ch. 14) used a different kind of manipulation of incentive reward factors in a film to teach pursuit-rotor tracking. Items nearer to a "reward sequence"(demonstrating the beneficial effect of a particular instructional point) were learned significantly better than those more distant from the reward sequence. It seems possible, however, that these findings are more closely related to the phenomena of attention-directing (cf. M & L Ch. 6) than to generalized motivational effects.

In one investigation of the effect of films on subsequent student activity, (M & L Ch. 13) no evidence was found that showing a film version of a novel induced pupils to read the book or even to withdraw it from the classroom library. However, May and Jenkinson (1953, M & L Ch. 14) later demonstrated that showing interesting but incomplete filmed episodes from a novel did significantly increase a number of pupils who withdrew the book from the library, though there was less clear evidence that many of them then read it and remembered the content. Several experiments have dealt with attempts to increase learning through general raising of drive by producing anxiety. An interesting example is the experiment by Feshbach and Janis (1951) who used several forms of instruction on care of the teeth. Three groups of subjects were given anxiety-producing communications such as close-up shots of decaying teeth and diseased gums; with degrees of anxiety-stimulating material, designed to evoke high, medium and mild degrees of anxiety. One might expect that the higher the anxiety drive aroused, the greater would be the learning. However, this was not found; rather, a lesser degree of anxiety was more effective. This result may be interpreted in terms of the hypothesis that the need of the high-anxiety group to reduce their anxiety led their attention to away from thinking about needs for care of the teeth, whereas a lesser degree of anxiety was stimulating attention and learning of the desired habits without being so aversive as to make subjects avoid thought that could lead to desired habits.

Another recent study in which interest effects of a program were studied in terms of subsequent behavior stimulated is that reported by

Gropper and Lumsdaine (1961X), who compared TV science demonstrations performed by an adult and by children, and found that having children do the demonstration (on the TV program) was more effective, as compared with having them done by the adult teacher, in terms of stimulating children in the viewing school classed to undertake similar experiments.

# Chapter 15

# Some Special Problems In Instructional-Media Research

**In this chapter several common problems in instructional media research are selected for special attention. Some of these topics are treated elsewhere in the guidebook, but they are singled out for discussion here because of the frequency with which these problems arise in instructional-media research.**

## Sampling of subject matter and the base for generalization of results.

Brunswick (1947) is noted for having called attention to the importance, in psychological or educational experiments, of having not only an adequately representative sampling of people but also of having an adequate sampling of materials or subject matter. The basis for this point is the danger of unsound generalization of results from experiments using only a single restricted set of materials when the conclusions may not apply to other subject matter or materials.

Granted the contention that some sampling of materials is always desirable, we are faced with the practical problem that it is often much harder and more costly to sample a wide range of material than it is merely to increase the size or range of individuals tested in an experiment. It is therefore well to recognize that the needs and importance of a sampling of subject matter may differ depending on the experiment and that replication of findings with respect to subject matter may be sought in several ways, some of which are less costly and difficult than others. However, the sampling of subject matter is in reality just one aspect of the more general requirement that the basis for generalization of any finding depends on the number of independent observations on which it is based.

Independent in what respects? From the standpoint of maximum generalizability of results, the ideal experiment on educational media would provide an explicit basis for generalization in which N, the number of observations on which experimental inference is based, would represent N completely independent observations, alike only in that they provided instances of the operation of the hypothesis that was under test. This condition would be very well approximated if a sign test were applied to the outcomes of N experimental tests of an hypothesis, each conducted with different subjects, a different subject matter, and different experimenters in different locations, possibly using different experimental designs. Under these ideal conditions we could be reasonably sure that rejection of a null hypothesis did not result from experimental artifacts, peculiarities of the specific sampling of subjects or learning materials. Somewhat less formal assurance that this is the case is provided when several more or less independent replications of an experiment are conducted, each of which separately obtains significant differences with respect to a particular factor commonly manipulated in the experiments.

Within any one experiment the secureness and breadth of the basis for generalizing the findings can in a sense be assessed from the degree to which N, in the test of significance employed, is based on observations which are independent with respect to as many relevant factors as possible. In rare instances, such as the experiments by Sheffield (1946) and Lumsdaine (1949, 1958), degrees of freedom can be based on observations each of which represented an independent sampling of learners each learning an independent sample of material drawn from some more or less well defined populations of learners and materials. Unfortunately, this is generally not done, and N is most commonly based on the number of independently assigned learners--or, in some cases (cf. pp. ), independently assigned groups of learners. Here, sampling of subject matter or material is rarely taken explicitly into account, although for some kinds of experimental designs it is more appropriate to base degrees of freedom on number of subject-matter items than on number of individual learners tested. For example (cf. Lumsdaine, 1958), if a design is employed in which each subject (learner) "serves as his own control" by learning one set of material under Experimental Condition I and a different set of material under Experimental Condition II, valid comparisons of I and II must take the sampling of subject-matter items explicitly into account in estimating error, since it is obvious that no matter how large the number of subjects, differences between I and II could depend solely on difference in the sample of material assigned to each; and no significance test not basing N on the number of items sampled could assess the likelihood that a given difference could have arisen by accident of differences arising from item sampling.

But in most experiments, subject matter is "controlled"--i.e., content is presumably the same for the two or more experimental treatments (Experimental Conditions.) In these instances, even though we should ideally proceed as above and base df on the number of independent replications of subjects and materials, we generally settle for basing df instead only on sampling of persons, leaving the generalizability to new subject matter up to faith or at most replicating with one or two subject matters to provide one or two df on which to satisfy ourselves as to this generalizability. Some further considerations with respect to replication are discussed below.

## Group versus individual testing with verbal and pictorial test materials

The question of whether to administer criterion tests (to assess the effects of instruction) individually or in groups is often confounded in practice with the related question, discussed above, of group versus individual sampling or assignment to experimental treatments. The same arguments apply to favor individual assignment to groups, and individual administration of an experiment often carries with it individual testing of subjects' achievement. However, the questions of group versus individual sampling, experimental instruction, testing, and application of results are by no means synonymous, and in theory one might employ any one of the 16 possible combinations of four dichotomous alternatives: (1) sampling by individuals versus sampling by pre-formed groups, (2) administration of the experimental instruments being compared in a group versus in an individual by (3) testing achievement in a group or individually, and (4) [redacted] the conclusions to later instruction that is to be presented in a group or in an individual situation. For example, in the Hovland, Lumsdaine and Sheffield study of active versus passive review (1949, Chap. 9), both group and individual testing were employed, though sampling was on an individual basis and experimental administration of instructional treatments was done in groups. Here, however, the applicability of findings favoring active review to an individual-instruction situation was limited by the factor that the active review procedure used employed oral group responses that provided inter-person feedback. On the other hand, the general principles favoring active response, as exemplified in written-response procedures used in later experiments such as that of Michael and Maccoby (1953, 1961), would also be applicable to individual instruction using teaching machines or other devices.

It is also sometimes the case that, regardless of whether sampling, experimental instruction, or desired application is by individual or by group, individual testing may have advantages with respect to sensitivity, reliability or demonstrability of consistency in the results. For example, in the above cited experiment on active review by Hovland, Lumsdaine and Sheffield (1949, Ch. 9), individual-test results were found to be more sensitive and useful for most of the analyses than were group test results. The individually administered, one-item-at-a-time tests controlled the amount of time devoted to each item in a way that group-administered paper and pencil tests did not, and also allowed measurement of individual reaction time. The latter permitted examining results in terms of correct responses given promptly as compared with those given more hesitantly (cf. Figure 3, Hovland, Lumsdaine and Sheffield, p. 235). On the other hand, the group tests provided a more economical source of data for analyses of differential effects on population subgroups. In a later replication of this experiment by Lumsdaine and Gladstone (1958), the advantage of item-exposure control (though not the availability of reaction time data) was realized in combination with the economy of group administration through the use of projected slides for group presentation of test items one at a time for a predetermined interval. (The latter could, of course, be varied as desired from item to item.)

Another useful aspect of the use of optically projected visual test materials is seen in the study of animation technique (Lumsdaine, Sulzer, & Kopstein (1961) in which criterion test performance was closely related to the criterion behavior which the film was designed primarily to teach--namely, the reading of micrometer setting. Although actual performance tests with real micrometers were not employed, the effective stimulus that would be encountered in such an actual performance test was realistically simulated by means of slides, each showing micrometer setting which the trainees were required to read. Fifteen different settings were used in the test, given after the exposure to any one of the experimental films. (In addition, other test questions were used to test effects on knowledge of nomenclature construction principles of operation of the micrometer.) In testing ability to read micrometer settings, each slide was projected on a screen for a fixed interval of time, and the trainees were instructed simply to write down the value of the settings on a test blank. [redacted] This test had a high face validity, and furthermore preliminary experimentation [redacted] indicated that if men could read the pictured settings they could also read the settings given in actual micrometers to read.

An interesting variant of this procedure devised by the writer was used in connection with a later study described in an unpublished report by Kanner and Sulzer (1961), in which objective scoring was employed without eliminating the advantage of having constructed, rather than recognition, responses by the examinees. This was done by presenting a projected picture of a micrometer setting for an appreciable period of time, during which the examinee had to figure out the reading and write it down. Then, for a brief interval of time, four possible numerical values for the setting were flashed on the screen, for a sufficiently brief interval that the examinee had time only to identify which (if any) of the four possible settings matched the one he had "constructed" (figured out) for himself. The four values were arranged horizontally coupled with "boxes" that simulated the corresponding spaces of the IBM answer sheet on which the student recorded his answer by indicating whether his answer for that question matched the value labled a, b, c, or d (or none of these.)

## Time as an independent and dependent variable in comparing instructional treatments

A special problem in experimental rationale is that of the interpretation to be placed on the expenditure of instructional time. This has been commented on with respect to the studies of Coulson and Silberman (1959), Fry (1960), and Evans, Glaser and Homme (1959). The most immediately obvious specific dilemma is that posed when results are reported, such as those of Coulson and Silberman, in which a measure of time spent in learning favors one group, whereas a measure of achievement favors the other group. Such troublesom problems of interpretation do not, of course, arise directly so long as experimentation is confined to measuring the comparative gains in achievement that result from fixed-pace instructional instruments in which total time is "controlled" at the same value for both of the alternate presentations.

There are, however, two more general problems concerning time which, though not peculiar to auto-instructional media, are brought into sharper focus because of the growing interest in such media and their potential capability for adjusting time spent to the capabilities of the learner. The first of these problems is the question of whether the relevant question to ask concerning the effectiveness of any form of instruction is, on the one hand, how high a score (less than perfect) is achieved by the end of a fixed, pre-set period of instruction (the length of which is often fixed by arbitrary administrative and scheduling considerations), or [redacted] to ask [redacted] how long it will take, by one form of instruction or another, to get students up to some pre-set minimum level of desired proficiency. Though there is no simple answer to this question, it needs at least to be identified as one which must be faced as not only interacting intimately with the overall philosophy and organization of education, but as having quite drastic implications for the techniques of assessing instruction and for the design of experimental assessment procedures as well as the design of instructional media themselves. How, for example, do we design the test sequences of an auto-instructional program to provide a valid indicator that criterion level has been reached for a particular portion of the subject matter--especially when an implicit ingredient of the criterion is to predict that the capabilities achieved will also be retained over an extended period following the completion of the program?

The second and related problem is one that presages some of the difficulties to be discussed in the following section with respect to the standards of sensitivity which an experiment ought to meet. This is the question of how to determine or even estimate reliably the "payoff" value of achievement scores as compared with time expenditure devoted to instruction. Use of some relative measure such as amount learned per unit time (see Goldbeck, 1960) might for a while help us duck the question, but its fundamental importance will persist. What are the relevant bases for deciding on the instrument method of choice, one of which achieves a higher achievement at the cost of more instructional time: there seems no way to answer this highly pertinent question unless we arrive at some defensible, uniform way of assessing the utility of instruction.

3. "Replicative" criterion measures

Experimental findings may, be called into question because of the arbitrariness of selection of a particular criterion variable. For example, in studies of attitude change (e.g., Hovland, Lumsdaine and Sheffield, Chapter 2) it may happen that some criterion questions show significant effects and others do not. As a related finding, in the data for a World War II experiment on persuasive effects (on men's estimates of how much longer the war would last), presenting one side versus both sides of a controversial question (Hovland, Lumsdaine and Sheffield, Chapter 8), "internal-change" measures indicating the net percentage who changed their opinions to any extent furnished a considerably more sensitive indicator than did percentages based on shifts in marginal distributions or changes in mean values of men's estimates. The same may be true of test items used to measure instructional effects. Furthermore, of the several ways in which criterion measures may be derived from such items, some may be sensitive to differential effects of alternative treatments and others may be less so. Consequently, confidence in the conclusion that a result does not depend on peculiarities of a particular measure, as well as confidence in the fact that an experiment is sensitive in revealing any effects that took place, may be increased by examining separately each of several relevant independent indicators, or even each of several ways of summarizing the data. As an illustration, in the earlier cited study of animation devices by Lumsdaine, Sulzer & Kopstein (1951) the experimental results were examined in terms of several criteria deriving from the same basic trainee performance. For example, mean scores (expressed as mean per cent of correct reading) showed a mean of 56.3% for the animated version as opposed to 47.4% correct for the unanimated version. In addition, the relative effectiveness of the two versions was also observed by comparing the results for the two groups in terms of the percentages of men who achieved various criteria of measured proficiency. For example, 72% of the men who saw the animated film answered at least three test questions correctly but only 57% of the nonanimated film groups were unable to do so. Similarly, 59% of the animated group got more than half the test questions correct, while only 56% of the nonanimated group did this well. The results in this case were thus consistent in direction, though with variation in the absolute magnitude of "effects" for each of several possible means of expressing the results. A further advantage of showing results for several different measures in experimental studies is that the specific numerical values used to indicate the "amount of superiority" for one experimental version over another will vary sometimes widely, depending on how the results were expressed. In some instances, indeed, different conclusions may be arrived at depending on which measure is selected as the criterion. Thus, conservatism is served by examining several different measures. When, as in this case, all of the measures lead to the same conclusion, the experimenter can rightly claim more confidence for his conclusions than when they are based only on a single measure of effect.

### "Hunting."

However, though it is often possible in an experiment to analyze the data in several different ways, it should be noted that there is a trap here: if one does analyses of the same data, and has set the 5% level of significance (rejection) as acceptable, he can expect to find a significant difference among or between treatments one time out of the twenty simply by chance. This is the theoretical reason for doing an analysis of variance before doing t-ratios for many different possible comparisons in a complex experiment. Doing so increases the likelihood of rejecting the null hypothesis when, in fact, the differences were simply due to random fluctuations in the data. On the other hand, as noted in Chapter 11, there may be several different ways of handling the data in analysis--some of which may be more sensitive than others--and there may be several different kinds of analysis which are pertinent to the same data.

In general it is perfectly legitimate to make any analysis of the data that seems sensible to E, provided that he then states clearly and precisely what it was that he did.

**Regression.**

Appendix D. of Hovland, Lumsdaine, and Sheffield (1949) is devoted to a problem that is of special importance in the testing of hypotheses which predict effects of instructional or communication media as a function of initial characteristics of the learner or audience. This is the tricky and often-ignored problem of regression, due to unreliability of the measuring instrument, as it enters into the analysis of the effects of a communication. This discussion might well be considered as "must" reading for any researcher who proposes to test hypotheses which predict effects in opposite directions for those differing in terms of initial response or in terms of characteristics correlated with initial response. It demonstrates that, in general, a valid analysis for testing such hypotheses must include paired before-after measures not only for the experimental groups but also for a control group which is needed to serve as an index of the amount of downward change from higher scores and upward change from lower ones which are inevitable in direction but generally unpredictable in extent in the case of all repeated observations with measuring instruments of less than perfect reliability. The importance of regression has been pointed out by HLS in the following terms:

"The tendency of extreme scores to regress toward the average is of great importance when one is analyzing the effects of an experimental variable on those with varying initial positions, because of the artifact that is introduced if the phenomenon is not controlled. A good example of this artifact

**has been pointed out by McNemar in his criticism of studies of the effect of environment on I.Q. In these studies children with low initial I.Q. scores were given special training and then retested. An average improvement in retest score was found, but it was subsequently pointed out by McNemar that improvement would be expected in this initially low-scoring group simply because of regression due to the presence of a certain amount of unreliability in I.Q. measurement..."**

**For further discussion of this topic, the reader is again referred to Ibid., Appendix D.**

### 3. The null hypothesis and the interpretation of "negative results"

It is well known that the logic of statistical inference underlying most experiments on instructional media and educational methods, as well as experimentation in the behavioral and biological sciences more generally, is one in which the data are analyzed with reference to a "null hypothesis" which asserts that two or more samples of data are in fact random samples from a common population. That is, the null hypothesis supposes that there are no differences between the data that result from two or more experimental treatments (e.g., two forms of instructional presentations), and the experimenter then proceeds to test the feasibility of this hypothesis. He does so by reference to appropriate tabulations or calculations which indicate the probability that, _if_ the null hypothesis were in fact ture, differences as large and/or as consistent as those he observes could occur simply as a consequence of random sampling fluctuations. If this probability is lower than some value (such as .01) agreed on conventionally as an acceptable level of statistical "significance," he rejects the null hypothesis. In doing so, he concludes that it is preferable to accept the alternative conclusion that the observed difference resulted from some non-random causal factor. This factor, in a well controlled experiment, he will normally indentify as the difference in the experimental treatments he employed.

There are many elaborations and sophisticated statistical details in the application of this null-hypothesis rationale, but all of them involve the basic argument just stated. When "positive" results are obtained--i.e., when the data are shown, as above, to be "incompatible" with a null hypothesis of no difference--the interpretation generally seems clear cut, and the experimenter rightly regards his data as having shown that a "significant difference" is present. That is, he has a convincing case for claiming that a non-chance difference has been demonstrated, since his statistical test has shown that it is very unlikely that his data would result if no systematic difference were present.

But difficulty arises when the data do not afford a basis for rejecting the null hypothesis at a level of confidence which has been agreed on as significant. The root of the trouble in this case is that there is then no logically defensible basis for drawing any conclusion about the true state of affairs. As R. A. Fisher long ago unequivocally pointed out, the null hypothesis can only be disproved; it cannot be proved. This is because any set of data compatible with a null hypothesis of no difference are also compatible with a number of alternative hypotheses that some difference does exist, even though it is not gross enough, in relation to the variability of the data, to be established as significant. Logically therefore, failure to disprove the null hypothesis leaves the experimenter (and those who wish to interpret or utilize his results) no defensible position except that of suspended judgment. This, of course, is not a position likely to be congenial to a person who has to make a practical decision, nor is it often congenial even to the experimenter who has toiled diligently in the hope of arriving at some definite conclusion. The temptation is therefore great to translate the only defensible statement of findings--namely, that the results merely fail to show evidence that there was a significant difference--into some more equivocal and more palatable form--e.g., that "no significant difference was found to exist," or that "there is no significant difference between the treatments" (with a semantic shift on the meaning of significant), or simply that "the two treatments did not differ significantly" (with the last word as often as not deleted in the final succinct summary written for the man who just wants the results and doesn't care how they were arrived at.)

That this kind of often unintentional verbal legerdemain which translates actually inconclusive results into definitive-sounding "negative results" (that are then cited to show that a particular factor is unimportant or "makes no difference") is denounced by logicians seems to have little effect on those who are wont to accept such indefensible but definite-sounding negative conclusions. Aside from wish-fulfillment, ignorance, or plain sloppy thinking, there are perhaps two reasons why such unsupportable conclusions are not more often successfully challenged. The first reason is that plainly there must be variables that are relatively unimportant in influencing any outcome, and it would often be as helpful to identify these as it is to identify those variables which do have an important influence. The second reason is that in an extended series of experiments on a number of relevant factors, we should expect that by and large the more important or more potent sources of variation would be the ones that would usually show up as statistically significant differences, and that by and large the unimportant or weak sources of variation would not. But this expectation will be borne out if and only if one crucial condition is fulfilled--namely, that the experiments are of comparable sensitivity, so that equally potent factors in the different experiments have about the same chance of showing up as significant. The fact that this condition is in general not fulfilled over the range of educational experiments creates a morass of ambiguity and inconsistency which has led some (including the writer) to wonder whether it is worth doing experiments at all until some basis for achieving comparable sensitivity from experiment to experiment is achieved.

The basis for this grave concern can perhaps be seen more clearly by considering four hypothetical (but perhaps not atypical) experimental situations. Experiment I sets out to test the influence of a variable which, in the knowledge of an omniscient Deity, is of finite though trivial importance (though the experimenter doesn't know this.) The experiment is, however, executed with great precision--because optimally sensitive measures are employed, because large numbers of cases are tested, and/or because extraneous sources of variation are under tight procedural control. The difference between treatments is found, though small, to be highly significant. Experiment II is concerned with assessing the effects of another variable, also of small actual potency, but is conducted with less precision, and leads to a conclusion of no significant difference. Experiments III and IV exactly parallel Experiments I and II, respectively with respect to their precision and their outcomes, but they have dealt with experimental variables which, the Deity knows, are in fact important, potent sources of variation. (This fact does not show up as a significant difference in Experiment IV, however, because of the insensitivity of the experiment.) The upshot is that, when the results are duly reported, the determination of which factors "made a significant difference" and of those which "had no significant effect" depends solely on the differential sensitivity of the several experiments, and not at all on the real importance of the factors that were investigated.

Now, of course, the situation as it exists over the range of actual experiments conducted on factors in instructional media (or on other factors studied in psychological and educational research) is not actually quite this bad. But a close look at the literature suggests that it may well be nearly this bad, and the point is that, in the absence of suitable, uniformly applicable criteria of experimental sensitivity, it is nearly impossible to know just how bad it is. Certainly in many a reported experiment the absence of significant differences is readily attributable to the insensitivity (due to one or more causes) of the experimental situation, and certainly in some instances (perhaps less frequent) attention has been unduly focused on a factor found to produce significant effects because of its having been studied in an experimental situation possessed of what we are generally tempted to think of as being commendably high precision or sensitivity. It is clear that the dismal hypothetical picture outlined above is certainly to some degree approximated, and will continue to be, until either (a) some agreed upon way is found for setting standards of experimental sensitivity that take into account all of the main sources of differential sensitivity, or (b) we are able to abandon the null-hypothesis paradigm as such, perhaps in favor of a paradigm yielding a workable estimation statistic which, in effect, can indicate, within specifiable limits of error, the proportion of total relevant variation in attainment (loosely speaking) which is attributable to each factor studied experimentally. The attaining of even a rough working approximation to a real solution for the dilemma in research interpretation posed by the undetermined-sensitivity dilemma is likely to be exceedingly difficult. But the author believes that an attempt to progress forward the explication of this basic methodological dilemma should be accorded AA-1 priority.

Pending basic progress in this respect, some possible recommendations to make are these: (1) that experimenters address themselves primarily to studying the interactions of factors they strongly suspect *will* produce significant differences in experiments of reasonable precision; (2) that they try to design their experiments to incorporate at least the available elementary safeguard of incorporating one or more factors which it is known in advance should make a difference (see Ch. 11, p ... ), as a basis for defending that at least minimal sensitivity has been realized; (3) that when non-significant differences are obtained they refrain religiously from stating their findings in ways that may be construed as other than indicating that the findings are inconclusive; (4) that all data and circumstances which might reasonably be found relevant to an assessment of experimental sensitivity be reported in full; and (5) that a determined effort be made to impress on utilizers and interpreters of research findings the limitations imposed by current methodological lacunae on the interpretations which can defensibly be made from experimental results. These limitations should include the fact that at present we (1) *can* establish (at any agreed-upon level of confidence) that a difference exists; (2) *may*, however, claim this erroneously if already known methodological safeguards (e.g., see the above discussion on group sampling) are not well complied with; (3) have very little basis for consistant statements about the magnitude, potency or importance of obtained significant differences; and (4) have no way at present to establish flatly, by any consistent standards, that a given factor does not make a difference (e.g., to show that two educational procedures or instruments are "equally effective.") It seems obvious that merely to state the last stricture is sufficient to emphasize (in view of the obvious practical need to be able to make such assessments) the importance of methodological advances beyond the customary usage of the null hypothesis rationale.

## Possible solutions to the methodological dilemma

### Decision-theory approaches

One approach to the partial solution of the problem of interpreting negative results in experiments is, of course, the use of fiducial limits to permit a conclusion that the difference, if any, resulting from the two methods is no greater than some specified amount. Though this approach is straightforward, it is relatively seldom used, and suffers from the difficulty that it does not help with the basic problem of deciding what the "specified amount" ought to be (see below.) A more thoroughgoing approach would seem to be to formulate questions of the comparative effectiveness of educational media in terms of a decision-theory rationale. Cronbach and Gleser (1957) have discussed the applicability of decision-theory paradigms to problems of personnel selection, and W. Edwards (1956) has discussed the theory of statistical decision functions (cf. Chernoff and Moses, 1959; Wald, 1950) in deciding on the method of choice as between two training methods. Edwards presents a very interesting hypothetical example in which calculations based on an appropriate model may be used to decide which of two competing methods is economically preferable. The example is worked out in such a way as to identify requirements that would have to be met to make utilization of this approach possible in a practical experimental situation. Perhaps the most significant aspect of his analysis is that it brings one face to face with our general inability to make reasonable estimates of the costs and, particularly, the anticipated payoff, from educational procedures. The ability to do so, directly or indirectly, would appear to be a crucial requirement for any decision-theory solution to the problem. Hence, while decision theory approaches may well prove ultimately to provide the best experimental rationale for the comparison of alternative educational instruments or procedures, a prior task would seem to be the exploration of ways to estimate educational costs and payoffs and, perhaps as a first step, the question of whether it is at least possible to arrive at agreement as to what constitutes an educationally worthwhile difference in the effectiveness of two compared alternatives. The latter question can also be discussed *per se*, without explicit reference to its utilization in decision-function calculations. Some aspects of the problems involved can in fact be perceived in relation to their interim application in attempting to set standards of sensitivity for experiments following a null hypothesis rationale, as discussed briefly in the following section.

## Improving the null hypothesis

If we are not yet in a position to supercede the null hypothesis rationale for experiments, and since the null hypothesis cannot be *proved*, we may ask how, as an interim measure at least, its application to the design and conduct of educational experiments can be *improved*. How, that is, can we agree on standards for the sensitivity of differences and the decision as to which magnitudes of effects shall consistently be regarded as negligible and which as educationally noteworthy, so as at least to improve somewhat the confusion and inconsistency which presently attends the interpretation of experimental findings? The following tentative suggestions might be offered in addition to the experimental and reportorial recommendations suggested above on pages 170, 171, as possibly susceptible of improving the current discouraging situation. Of necessity, these suggestions are tentative, incompletely formulated, and therefore offered not without hesitation. It is the author's belief, however, that at least some of them may afford one basis for our beginning to make some much needed improvements in the state of the art in educational experimentation. At a

15-18

minimum it would seem defensible that such possibilities ought at least to be considered by the prospective experimenter who seeks to discover laws and principles concerning the operation of factors that regulate the effectiveness of instructional media.

A first step that can apply in any experiment is to decide in advance how small a difference in the criterion measure the experimenter is willing to regard as of negligible importance, and then to design his experiment so that any actual differences larger than this value will, if present, be likely to show up as significant. A clear, unequivocal initial statement of where this just detectable difference has been set would in some cases lead to a decision not to do the experiment; in other cases it would lead the experimenter to increase the sensitivity of his experiment (perhaps in the interests of economy, sometimes through a sequential set of replications); in all cases it would probably clarify the interpretation to be placed on negative results, fostering statements such as "the difference if any is less than such and such" rather than "no difference exists." Insistence by editors of technical publications on reportorial scrupulousness in this respect might help; further, it would be worth trying to see if standardized phraseologies for different kinds of experimental outcomes might not be agreed on and enforced.

A second step that might be useful in extended programmatic research would be the adoption of a few standard subject matters, coupled with criterion tests of known, consistent properties. Several such subject matters might be chosen to represent various classes of learning problems (e.g., a standard set of language materials, a standard science subject, a standard motor-skill task.) Despite some obvious disadvantages in doing this, the practice would tend to minimize the extent to which results were governed by subject matter-linked differences in experimental sensitivity. Use of a rather restricted set of standard experimental subject matters would also facilitate the arriving at agreement on an acceptable value for just-detectable differences.

Another probably useful practice might be to discourage the reportorial convention of merely dichotomizing results as significant or nonsignificant, and encourage the reporting of actual probability levels obtained (stipulating in each case whether one-tail or two-tail values are being reported.) Doing so would not only call attention to the fact that experimental results actually represent varying degrees of confidence but might also encourage suspended judgment and (hopefully) the tencency to replicate "borderline-significance" results.

A fourth helpful tendency might be brought automatically with an increased tendency to study contingent propositions rather than global assessments and to introduce co-varied experimental factors previously found to be effective (see page ) It is, of course, possible to demonstrate that one difference is less than another even though one cannot demonstrate that the first is zero.

Probably the most difficult and fundamental direction of progress, however, is to develop rationales more satisfactory than an individual experimenter's purely arbitrary decision for deciding on the magnitude of a just-detectable difference. Several possibilities, all of them far from fully satisfactory, may be suggested.

First: we might try to select experimental subject matter in terms of criterion variables that seem to be the relatively more tangible ones in terms of immediate educational payoff. Experiments on instruction used in vocational training are perhaps closest to filling the bill.

Second: as one means to this end, we might try to increase the proportion of experiments in which time to reach a fixed criterion of proficiency is the criterion variable, rather than scores arrived at in a fixed time interval. Difficult as it may be to translate instructional time into economic payoff (at least in general education) it is probably less difficult than to make this translation for differences in test scores.

Third: we might make a uniform practice of trying to obtain estimates from at least several presumably qualified judges as to how large a difference in test scores (or in instructional time) can reasonably be left undetectable. At least this might

15-20

give a reliable basis, if one of doubtful validity, that would be preferable to purely arbitrary decision on the part of one experimenter. We might also try to see if more meaningful judgments would result if the just-detectable difference were expressed in terms of grade norms (e.g., the experiment should at least be able to detect as significant the difference between an "A" and a "B"), or age norms (the experiment might be regarded as unsatisfactorily insensitive if incapable of detecting the difference between average fifth and seventh grade pupils.)

Fourth: the last proceeding might be more meaningful if test-scores differences were equated to instructional-time differences: for how large a test-score difference would we be willing to increase instructional time by 10%?

Fifth: we could try to agree on minimal detectable differences in relation to population variation in the criterion measure, perhaps agreeing that no experiment should be conducted that will probably not detect as significant a difference of say, 1/2 S.D. in the distribution of some standard student population.

Sixth: for at least a considerable range of subject matters, we might similarly specify minimal satisfactory sensitivity as the ability of an experiment to reveal differences associated with prediction variables--e.g., that the experiment should be capable of detecting a difference as large as that between students differing in 15 I.Q. points.

Finally, and most generally, we might--and, perhaps must--make a determined effort to see what degree of agreement can be obtained by these or, hopefully, better approaches on some standard for setting the sensitivity of our experiments, whereby we can begin to report on a more consistent, describable basis the grounds for paying more attention to "significant" than to "non-significant" experimental results.

# BIBLIOGRAPHY[1, 2]

Allen, W. H. Readability of instructional film commentary. *J. appl. Psychol.*, 1952, *36*, 164-168.

Allen, W. H. Audio-visual materials. *Rev. educ. Res.*, 1956, *26*, 125-156.

Allen, W. H. Research on film use: student participation. *Audio-Visual Com. Rev.*, 1957, *5*, 423-450.

Allen, W. H. Audio-visual communication research. In C. W. Harris (Ed.), *Encyclopedia of educational research*. (3rd ed.) New York: MacMillan Co., 1960. Pp. 115-137.

Anderson, N., Brainerd, R. C., & Rath, G. The IBM Research Center teaching machines project. Yorktown Heights, N. Y.: IBM Research Center, unpublished manuscript, 1958.

Angell, D., & Lumsdaine, A. A. *Prompted plus unprompted trials versus prompted trials alone in paired-associate learning*. Pittsburgh, Pa.: American Institute for Research, 1960. (See also Angell & Lumsdaine, 1961).

L Angell, D., & Lumsdaine, A. A. Prompted and unprompted trials versus prompted trials only in paired-associate learning. (Ch. 24) 1961.

Angell, G. W. Effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. *J. educ. Res.*, 1949, *42*, 391-394.

LG Angell, G. W., & Troyer, M. E. A new self-scoring test device for improving instruction. *Sch. & Soc.*, 1948, *67*, 84-85. [66-68].

---

[1] All references preceded by "LG" are reprinted in A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning: a source book*. Page numbers in brackets refer to this volume. "TMPL abstr." indicates that this volume contains an abstract of the study after which the notation appears.

[2] All references preceded by "L" are chapters in A. A. Lumsdaine (Ed.), *Student response in programmed instruction: a symposium*. Chapter numbers are in parentheses.

Ash, P. *The relative effectiveness of massed versus spaced film presentation.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-3, 1949.

Ash, P., & Carlton, B. J. *The value of note-taking during film learning.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-21, November 1951.

Ash, P., & Jaspen, N. *Optimum physical viewing conditions for a rear projection daylight screen.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-37, 1953.

Aukes, L. E., & Simon, G. B. *The relative effectiveness of an Air Force training device used intact versus with isolated parts.* Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, Res. Rept. AFPTRC-TN-57-77 (ASTIA Doc. No. 131429), June 1957.

LG Briggs, L. J. Two self-instructional devices. *Psychol. Rep.*, 1958, *4*, 671-676. [299-304].

L Briggs, L. J. Prompting and confirmation conditions for three learning tasks employing the Subject-Matter Trainer. (Ch. 23) 1961(a).

L Briggs, L. J. Self-pacing versus automatic pacing in paired-associate learning. (Ch. 25) 1961(b).

Briggs, L. J., Plashinski, D., & Jones, D. L. Self-pacing versus automatic pacing of practice on the Subject-Matter Trainer. Air Force Personnel and Training Research Center, Unpublished Lab. Note ASPRL-LN-55-8 (mimeo.), Sept. 1955. (TMPL abstr., pp. 591-592).

Brunswik, E. *Systematic and representative design of psychological experiments, with results in physical and social perception*. Berkeley, Calif.: University of California Press, 1947.

Buckingham,

Carpenter, C. R. A challenge for research. *Educ. Screen*, 1948, *27*, 119-121.

Carpenter, C. R., & Greenhill, L. P. *An investigation of closed-circuit television for teaching university courses*. University Park, Pa.: Pennsylvania State University, Instructional Television Research Report No. 1, 1955.

Carpenter, C. R., & Greenhill, L. P. *An investigation of closed-circuit television for teaching university courses*. University Park, Pa.: Pennsylvania State University, Instructional Television Research Report No. 2, 1958.

Chall, Jeanne S., & Dial, H. E. Predicting listener understanding and interest in newscasts. *Educ. Res. Bull*., 1948, *27*, 141-153, 168.

Cheatham, P. G. *A comparison of the visual and auditory senses as possible channels for communication*. Wright-Patterson Air Force Base, Ohio: U. S. Air Force Air Materiel Command, A.F. Tech. Rept. No. 5919, May 1950.

Chernoff, H., & Moses, L. E. *Elementary decision theory*. New York: Wiley, 1959.

Cogswell, J. F. *Effects of a stereoscopic sound motion picture on the learning of a perceptual-motor task*. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-32, Sept. 1952.

Cook, J. O. Supplementary report: processes underlying learning a single paired-associate item. *J. exp. Psychol*., 1958, *56*, 455. (TMPL abstr., pp. 601-602.)

Cook, J. O. Research in audio-visual communication. In J. Ball & F. C. Byrnes (Eds.), *Research, principles, and practices in visual communication.* East Lansing, Michigan: Michigan State University, National Project in Agricultural Communications, 1960. Pp. 91-106.

L Cook, J. O. From audience participation to paired-associates learning. (Ch. 21) 1961(a).

L Cook, J. O. Response analysis in paired-associate learning experiments. (Ch. 22) 1961(b).

Cook, J. O., & Kendler, Tracy S. A theoretical model to explain some paired-associate learning data. In G. Finch & F. Cameron (Eds.), *Symposium on Air Force human engineering, personnel, and training research.* Washington, D. C.: National Academy of Sciences - National Research Council, 1956. Pp. 90-98.

Cook, J. O., & Spitzer, M. E. Supplementary report: prompting versus confirmation in paired-associate learning. *J. exp. Psychol.*, 1960, *59*, 275-276. (TMPL abstr., pp. 604-605.)

LG Coulson, J. E., & Silberman, H. F. *Results of an initial experiment in automated teaching.* Santa Monica, Calif.: System Development Corp., July 1959. [452-468].

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* Urbana, Ill.: University of Illinois Press, 1957.

Crowder, N. A. Automatic tutoring by means of intrinsic programming. In E. H. Galanter (Ed.), *Automatic teaching: the state of the art.* New York: Wiley, 1959. Pp. 109-116.

LG Crowder, N. A. Automatic tutoring by intrinsic programming. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning: a source book.* Washington, D. C.: National Education Association, 1960. Pp. 286-298.

Dale, E., & Chall, Jeanne S. A formula for predicting readability: instructions. Educ. Res. Bull., 1948, 27, 37-54.

Day, W. F., & Beach, Barbara R. A survey of the research literature comparing the visual and auditory presentation of information. Wright-Patterson Air Force Base, Ohio: U. S. Air Force Air Materiel Command, A. F. Tech. Rept. No. 5921, 1950.

Detambel, M. H., & Stolurow, L. M. Stimulus sequence and concept learning. J. exp. Psychol., 1956, 51, 34-40.

Duva, J. S., & Lumsdaine, A. A. The influence of image quality on the teaching effectiveness of the kinescope. U. S. Air Force Human Factors Operations Research Laboratories, Unpublished HFORL Report, March 1956.

Eastman Kodak Company. Legibility standards for projected material. Rochester, N. Y.: Kodak Pamphlet No. S-4, 1955.

Edwards, A. L. Statistical methods for the behavioral sciences. New York: Rinehart, 1954.

Edwards, W. The use of statistical decision functions in making practical decisions. In G. Finch & F. Cameron (Eds.), Symposium on Air Force human engineering, personnel, and training research. Washington, D. C.: National Academy of Sciences - National Research Council, 1956. Pp. 115-124.

Eigen, L. D., & Komoski, P. K. Research summary number 1: automated teaching project. New York: Collegiate School (mimeograph), June 1960.

Emeson, D. L., & Wulff, J. J. The relationship between "what is learned" and "how it's taught." Air Force Personnel and Training Research Center, Maintenance Laboratory, Tech. Memo. ML-TM-57-32, Dec. 1957. (See also Wulff & Emeson, 1961.)

Evans, J. L. An investigation of teaching machine variables using learning programs in symbolic logic. Unpublished doctoral dissertation, University of Pittsburgh, 1960.

LG Evans, J. L., Glaser, R., & Homme, L. E. A preliminary investigation of variation in the properties of verbal learning sequences of the "teaching machine" types. 1959. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning: a source book.* Washington, D. C.: National Education Association, 1960. Pp. 446-451.

Evans, J., Homme, L. E., & Glaser, R. *The Ruleg (Rule-Example) system for the construction of learning programs*. (A report prepared under the Cooperative Research Program of the United States Office of Education) Pittsburgh, Pa.: University of Pittsburgh, Dept. of Psychology (mimeograph), 1960. (TMPL abstr., 619-620.)

Fairbanks, G., Guttman, N., & Miron, M. S. Auditory comprehension in relation to listening rate and selective verbal redundancy. *J. Speech & Hearing Disorders,* 1957(a), 22, 23-32.

Fairbanks, G., Guttman, N., & Miron, M. S. Auditory comprehension of repeated high-speed messages. *J. Speech & Hearing Disorders,* 1957(b), 22, 20-22.

Fairbanks, G., Guttman, N., & Miron, M. S. Effects of time compression of connected speech. *J. Speech & Hearing Disorders*, 1957(c), 22, 10-19.

Faison, E. W. J., Rose, N., & Podell, J. E. A technique for measuring observable audience reactions to training films. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-45, (mimeo.) December 1955.

Feshbach, S., & Janis, I. L. An experimental study of the effects of anxiety-arousing appeals in mass communication. Paper read at Eastern Psychological Association, New York, March 1951.

Fisher, R. A. The design of experiments. (4th ed.) London: Oliver and Boyd, 1947.

Flesch, R. The art of plain talk. New York: Harper, 1946.

Flesch, R. The art of readable writing. New York: Harper, 1949.

LG Fry, E. B. A study of teaching machine response modes. In A. A. Lumsdaine & R. Glaser (Eds.), Teaching machines and programmed learning: a source book. Washington, D. C.: National Education Association, 1960. Pp. 469-474.

Gagné, R. M. The effect of sequence of presentation of similar items on the learning of paired associates. J. exp. Psychol., 1950, 40, 61-73.

Galanter, E. H. (Ed.) Automatic teaching: the state of the art. New York: Wiley, 1959.

Garvey, W. D., & Henneman, R. H. Practical limits of speeded speech. Wright-Patterson Air Force Base, Ohio: U. S. Air Force Air Materiel Command, A. F. Tech. Rept. No. 5917, May 1950.

Gates, A. I. Recitation as a factor in memorizing. Arch. Psychol., 1917, 7 (40).

Gibson, J. J. (Ed.) Motion picture testing and research. (Army Air Forces Aviation Psychology Program) Washington, D. C.: Government Printing Office, Research Rept. No. 7, 1947.

Gilbert, T. F. Some recent attempts at the partial automation of teaching. A report to the Bell Telephone Laboratories (unpublished mimeo.), Sept. 1958. (TMPL abstr., pp. 635-636.)

LG Gilbert, T. F. On the relevance of laboratory investigation of learning to self-instructional programming. In A. A. Lumsdaine & R. Glaser (Eds.), Teaching machines and programmed learning: a source book. Washington, D. C.: National Education Association, 1960. Pp. 475-485.

Gladstone, A. The readability of the commentary. In M. A. May & A. A. Lumsdaine (Eds.), Learning from films. New Haven: Yale University Press, 1958. Pp. 46-57.

Glaser, R. (Ed.) Training research and education. Pittsburgh, Pa.: University of Pittsburgh Press, 1962.

Goldbeck, R. A. The effect of response mode and learning material difficulty on automated instruction. Technical Report No. 1. Santa Barbara, Calif.: American Institute for Research, 1960.

Greenhill, L. P. A study of the feasibility of local production of minimum cost sound motion pictures. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-48, 1955.

Gropper, G. L., & Lumsdaine, A. A. The use of student response to improve televised instruction: an overview. Report No. 7, Summary of six prior reports. Studies in Televised Instruction. Pittsburgh, Pa.: American Institute for Research, 1961.

Guthrie, E. R. The psychology of learning. New York: Harper, 1935. (Rev. ed., 1952.)

Hall, W. H. A study of three methods of teaching science with classroom films. Sch. Sci. & Math., 1936, 36.

Harby, S. F. Evaluation of a procedure for using daylight projection of film loops in teaching skills. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-25, 1952.

Hirsch, R. S. *The effects of knowledge of test results on learning of meaningful material.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-30, September 1952.

Hoban, C. F., Jr., & van Ormer, E. B. *Instructional film research, 1918-1950.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-19, 1950.

Hoehn, A. J., & Lumsdaine, A. A. *Design and use of job aids for communicating technical information.* Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, Tech. Rept. AFPTRC-TR-58-7 (ASTIA Doc. No. AD 152109), January 1958. (TMPL abstr., p. 639.)

Holland, J. G. A teaching machine program in psychology. In E. H. Galanter (Ed.), *Automatic teaching: the state of the art.* New York: Wiley, 1959. Pp. 69-82.

Holland, J. G. Program design and use. Paper read at the American Psychological Association Convention, Chicago, 1960(a).

LG Holland, J. G. Teaching machines: an application of principles from the laboratory. In *Proceedings of the Educational Testing Service invitational conference.* Symposium on "The impact of testing on the educational process," October 1959. Princeton, N. J.: Educational Testing Service, 1960(b). [215-228.]

LG Homme, L. E., & Glaser, R. Problems in programming verbal learning sequences. Paper read at symposium on Research Issues in the Study of Human Learning Raised by Developments in Automated Teaching Methods, American Psychological Association Convention, Cincinnati, Sept. 1959. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning: a source book.* Washington, D. C.: National Education Association, 1960. Pp. 486-496.

Hovland, C. I. Human learning and retention. In S. S. Stevens (Ed.), Handbook of experimental psychology. New York: Wiley, 1951. Pp. 613-689.

Hovland, C. I. (Ed.) The order of presentation in persuasion. New Haven: Yale University Press, 1957.

Hovland, C. I., Janis, I. L., & Kelley, H. H. Communication and persuasion: psychological studies of opinion change. New Haven: Yale University Press, 1953.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. Experiments on mass communication. Princeton: Princeton University Press, 1949.

Hovland, C. I., & Weiss, W. The influence of source credibility on communication effectiveness. Publ. Opin. quart., 1951, 15, 635-650.

Hurlock, E. B. An evaluation of certain incentives used in school work. J. educ. Psychol., 1925, 16, 145-159.

Irion, A. L., & Briggs, L. J. Learning task and mode of operation variables in use of the Subject-Matter Trainer. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, Tech. Rept. AFPTRC-TR-57-8 (ASTIA Doc. No. AD 134252), October 1957. (TMPL abstr., pp. 642-643.) (See Briggs, 1961a.)

Israel, M. L. Variably blurred prompting: I. methodology and application to the analysis of paired-associate learning. J. Psychol., 1960, 50, 43-52.

Jackson, R. Learning from kinescopes and films. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 20-TV-1, 1952.

Jaspen, N. Contribution of especially designed sound motion pictures to the learning of skills: assembly of the breech block of the 40mm. antiaircraft gun. Unpublished doctoral dissertation, Pennsylvania State University, 1949.

Jaspen, N. *Effects on training of experimental film variables, study II: verbalization, "how-it-works," nomenclature, audience participation, and succinct treatment.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-11, March 1950.

Judd, C. H. The relation of special training to general intelligence. *Educ. Rev.*, 1908, *36*, 28-42.

Kale, S. V., Grosslight, J. H., & McIntyre, C. J. *Exploratory studies in the use of pictures and sound for teaching foreign language vocabulary.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-53, August 1955.

Kanner, J. H. Army television research activities. *Audio-Visual Com. Rev.*, 1960, *8*, 74.

Kanner, J. H., & McClure, A. H. Varied versus identical repetition in filmed instruction on micrometer reading. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-56-11 (mimeo.), April 1956.

Kanner, J. H., & Sulzer, R. L. Overt and covert rehearsal of 50% versus 100% of the material in film learning. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Staff Memorandum TARL 56-12 (mimeo.), 1956. (See Kanner & Sulzer, 1961.)

L Kanner, J. H., & Sulzer, R. L. Overt and covert rehearsal of fifty per cent versus one hundred per cent of the material in film learning. (Ch. 28) 1961.

Kendall, M. G. Hiawatha designs an experiment. *Amer. Statistician*, 1959, *13*, 23-24.

Kendler, Tracy S., Cook, J. O., & Kendler, H. H. An investigation of the interacting effects of repetition and audience participation on learning from films. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL 56-2 (mimeo.), January 1956. (Paper presented at American Psychological Association, Cleveland, Ohio, September 1953; see also abstract in *Amer. Psychologist*, 1953, *8*, 378-379; and Cook, 1961a.)

Kendler, H. H., Kendler, Tracy S., Desiderato, A. H., & Cook, J. O. *An investigation of the effect of audio-visually induced anxiety of learning and performance.* New York: New York University, 1953.

Kimble, G. A. The effect of praise, reproof, and no incentive on the value of a training film. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-7 (mimeo.), 1955(a).

Kimble, G. A. The value of an interpolated test in increasing the effectiveness of a training film. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-10 (mimeo.), 1955(b).

Kimble, G. A., & Wulff, J. J. *Response guidance as a factor in the value of audience participation in training film instruction.* Washington, D. C.: U. S. Air Force Human Factors Operations Research Laboratory, HFORL Memo. Report No. 36, 1953. (See Kimble & Wulff, 1961b.)

L Kimble, G. A., & Wulff, J. J. The effectiveness of instruction in reading a scale as influenced by the relative amounts of demonstration and problem-solving practice. (Ch. 16) 1961(a).

L Kimble, G. A., & Wulff, J. J. "Response guidance" as a factor in the value of student participation in film instruction. (Ch. 15) 1961(b).

Klare, G. R., Mabry, J. E., & Gustafson, L. M. The relationship of patterning (underlining) to immediate retention and to acceptability of technical material. J. appl. Psychol., 1955(a), 39, 40-42.

Klare, G. R., Mabry, J. E., & Gustafson, L. M. The relationship of style difficulty to immediate retention and to acceptability of technical material. J. educ. Psychol., 1955(b), 46, 287-295.

Klare, G.R., Nichols, W. H., & Shuford, E. H. The relationship of typographic arrangement to the learning of technical training material. J. appl. Psychol., 1957, 41, 41-45.

Klare, G. R., Shuford, E. H., & Nichols, W. H. The relationship of style difficulty, practice, and ability to efficiency of reading and retention. J. appl. Psychol., 1957, 41, 222-226.

Klare, G. R., Shuford, E. H., & Nichols, W. H. The relationship of format organization to learning. Educ. Res. Bull., 1959, 38, 49-50.

Klaus, D. J., & Lumsdaine, A. A. An experimental field test of the value of self-tutoring materials in high school physics. An Interim Report of Progress and Preliminary Findings. Pittsburgh, Pa.: American Institute for Research (mimeo.), April 1960.

Kopstein, F. F., & Roshal, S. M. Learning foreign vocabulary from pictures versus words. Amer. Psychologist, 1954, 9, 407-408. (Abstract.) (See Kopstein & Roshal, 1961.)

Kopstein, F. F., & Roshal, S. M. Method of presenting word pairs as a factor in foreign vocabulary learning. Paper presented at the annual convention of the American Psychological Association, San Francisco, California, September 1955. (See abstract in Amer. Psychologist, 1955, 10, 354; and Kopstein & Roshal, 1961.)

L Kopstein, F. F., & Roshal, S. M. Verbal learning efficiency as influenced by the manipulation of representational response processes: pictorial-verbal and temporal contiguity factors. (Ch. 20) 1961.

Kurtz, A. K., Walter, J. S., & Brenner, H. R. The effects of inserted questions and statements on film learning (rapid mass learning.) (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-16, September 1950.

Kurtz, K. H., & Hovland, C. I. The effect of verbalization during observation of stimulus objects upon accuracy of recognition and recall. J. exp. Psychol., 1953, 45, 157-164.

Lashley, K. S., & Watson, J. B. A psychological study of motion pictures in relation to venereal disease campaigns. Washington, D. C.: U. S. Interdepartmental Social Hygiene Board, 1922.

Lathrop, C.W., Jr. Contributions of film instruction to learning from instructional films. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Progress Report No. 13, 1949.

Levine, S. The role of motivation in the effects of "active review" on learning from a factual film. Amer. Psychologist, 1953, 8, 388-389. (Abstract.) (See Maccoby, Michael, & Levine, 1961.)

Lindquist, E. F. Design and analysis of experiments in psychology and education. New York: Houghton Mifflin, 1953.

LG Little, J. K. Results of use of machines for testing and for drill, upon learning in educational psychology. J. exp. Educ., 1934, 3, 45-49. [59-65]

Los Angeles City School Districts. An experiment in instructional TV in the Los Angeles junior colleges. Los Angeles: Division of Extension and Higher Education, 1959.

Lumsdaine, A. A. Experimental research and the improvement of teaching films. Educ. Screen, 1947, 26, 254-255.

Lumsdaine, A. A. Ease of learning with pictorial and verbal symbols. Unpublished doctoral dissertation, Stanford University, 1949.

Lumsdaine, A. A. Audio-visual research in the U. S. Air Force. Audio-Visual Com. Rev., 1953, 1, 76-90.

Lumsdaine, A. A. Cue and response functions of pictures and words. In M. A. May and A. A. Lumsdaine (Eds.), Learning from films. New Haven: Yale University Press, 1958. Pp. 123-149.

Lumsdaine, A. A. Partial and more complete automation of teaching in group and individual learning situations. In E. H. Galanter (Ed.), Automatic teaching: the state of the art. New York: Wiley, 1959(a). Pp. 147-166.

Lumsdaine, A. A. Response cueing and "size-of-step" in automated learning programs. Paper read at Symposium on Research Issues in the Study of Human Learning Raised by Developments in Automated Teaching Methods, American Psychological Association Convention, Cincinnati, September 1959(b). Preliminary version of Lumsdaine 1960(b).

LG Lumsdaine, A. A. Teaching machines and self-instructional materials. Audio-Visual Com. Rev., 1959(c), 7, 163-172. (Reprinted in Lumsdaine & Glaser, 1960, under the title, "Teaching machines: an introductory overview," pp. 5-22.)

Lumsdaine, A. A. Graphic aids, models and mockups as tools for individual and classroom instruction. In G. Finch (Ed.), Symposium on educational and training media. Washington, D. C.: National Academy of Sciences - National Research Council, 1960(a). Pp. 69-107.

G Lumsdaine, A. A. Some issues concerning devices and programs for automated learning. In A. A. Lumsdaine and R. Glaser (Eds.), Teaching machines and programmed learning: a source book. Washington, D. C.: National Education Association, 1960(b). Pp. 517-539.

Lumsdaine, A. A. Some differences in approach to the programming in instruction. In J. P. Lysaught (Ed.), Programmed learning: evolving principles and industrial applications. Ann Arbor, Michigan: Foundation for Research on Human Behavior, 1961(a). Pp. 37-52.

Lumsdaine, A. A. (Ed.) Student response in programmed instruction: a symposium. Washington, D. C.: National Academy of Sciences - National Research Council, 1961(b).

Lumsdaine, A. A., & Gladstone, A. Overt practice and audio-visual embellishments. In M. A. May and A. A. Lumsdaine (Eds.), Learning from films. New Haven: Yale University Press, 1958. Pp. 58-71.

Lumsdaine, A. A., & Glaser, R. (Eds.) Teaching machines and programmed learning: a source book. Washington, D. C.: National Education Association, 1960.

Lumsdaine, A. A., & Janis, I. L. Resistance to counter-propaganda produced by one-sided and two-sided propaganda presentations. Publ. Opin. quart., 1953, 17, 311-318.

Lumsdaine, A. A., May, M. A., & Hadsell, R. S. Questions spliced into a film for motivation and pupil participation. In M. A. May & A. A. Lumsdaine (Eds.), Learning from films. New Haven: Yale University Press, 1958. Pp. 72-83.

Lumsdaine, A. A., & Roshal, S. M. (Eds.) Experimental research on educational media. Los Angeles: University of California, 1962.

Lumsdaine, A. A., & Sulzer, R. L. The influence of simple animation techniques on the value of a training film. Washington, D. C.: U. S. Air Force Human Factors Research Laboratory, HRRL Rept. No. 24, 1951. (See Lumsdaine, Sulzer, & Kopstein, 1961.)

L Lumsdaine, A. A., Sulzer, R. L., & Kopstein, F. F. The effect of animation cues and repetition of examples on learning from an instructional film. (Ch. 17) 1961.

L Maccoby, N., Michael, D. N., & Levine, S. Further studies of student-participation procedures in film instruction: review and preview covert practice, and motivational interactions. (Ch. 19) 1961.

Maccoby, N., & Sheffield, F. D. Theory and experimental research on the teaching of complex sequential procedures by alternate demonstration and practice. In G. Finch & F. Cameron (Eds.), Symposium on Air Force human engineering, personnel, and training research. Washington, D. C.: National Academy of Sciences - National Research Council, Publication No. 516, 1958. Pp. 99-107.

L Maccoby, N., & Sheffield, F. D. Combining practice with demonstration in teaching complex sequences: summary and interpretation. (Ch. 5) 1961.

L Margolius, G. J., & Sheffield, F. D. Optimum methods of combining practice with filmed demonstration in teaching complex response sequences: serial learning of a mechanical-assembly task. (Ch. 3) 1961.

Margolius, G. J., Sheffield, F. D., & Maccoby, N. *Repetitive versus continuous demonstration and practice in the learning of a serial mechanical assembly task I*. Air Force Personnel and Training Research Center, Maintenance Laboratory, Tech. Memo. ML-TM-57-11, September 1957(a). (See Margolius, Sheffield & Maccoby, 1961a.)

Margolius, G. J., Sheffield, F. D., & Maccoby, N. *Repetitive versus continuous demonstration and practice in the learning of a serial mechanical assembly task II*. Air Force Personnel and Training Research Center, Maintenance Laboratory, Tech. Memo. ML-TM-57-12, September 1957(b). (See Margolius, Sheffield & Maccoby, 1961a.)

Margolius, G. J., Sheffield, F. D., & Maccoby, N. *Timing of demonstration and overt practice as a function of task organization*. Fir Force Personnel and Training Research Center, Maintenance Laboratory, Tech. Memo. ML-TM-57-13, September 1957(c). (See Margolius, Sheffield & Maccoby, 1961b.)

L Margolius, G. J., Sheffield, F. D., & Maccoby, N. Repetitive versus consecutive demonstration and practice in the learning of a serial mechanical-assembly task. (Ch. 6) 1961(a).

L Margolius, G. J., Sheffield, F. D., & Maccoby, N. Timing of demonstration and overt practice as a function of task organization. (Ch. 7) 1961(b).

May, M. A. The psychology of learning from demonstration films. *J. educ. Psychol.*, 1946, *37*, 1-12.

May, M. A. Verbal responses to demonstrational films. In M. A. May & A. A. Lumsdaine (Eds.), *Learning from films*. New Haven: Yale University Press, 1958.

May, M. A., & Jenkinson, Nelle L. Developing interest in reading with film. *Audio-Visual Com. Rev.*, 1953, *1* (3), 159-166.

May, M. A., & Jenkinson, Nelle L. The reading of a book based on a film. In M. A. May & A. A. Lumsdaine (Eds.), *Learning from films*. New Haven: Yale University Press, 1958. Pp. 195-203.

May, M. A., & Lumsdaine, A. A. *Learning from films*. New Haven: Yale University Press, 1958.

McClure, A. H., & Peterman, J. N. Filmed instructions for marking machine-scored test forms. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-56-7 (mimeo.), Jan. 1956.

McCollough, Celeste, & Van Atta, L. The use of miniature programs to supplement conventional teaching techniques. Paper read at the American Psychological Association Convention, Chicago, 1960.

McGuire, W. J. The relative efficacy of overt and covert trainee participation with different speed of instruction. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-46 (mimeo.), December 1955(a). (See McGuire, 1961a.)

McGuire, W. J. Studies of factors influencing the effectiveness of demonstrational film for teaching a motor skill: II. slow motion, added narration and distributed showings. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-50 (mimeo.), December 1955(b). (See also McGuire, 1961c.)

McGuire, W. J. Studies of factors influencing the effectiveness of demonstrational films for teaching a motor skill: III. serial position and proximity to reward within the film. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-51 (mimeo.), December 1955(c). (See also McGuire, 1961b).

L McGuire, W. J. Audience participation and audio-visual instruction: overt-covert responding and rate of presentation. (Ch. 27) 1961(a).

L McGuire, W. J. Effects of serial position and proximity to "reward" within a demonstrational film. (Ch. 14) 1961(b).

L McGuire, W. J. Some factors influencing the effectiveness of demonstrational films: repetition of instructions, slow motion, distribution of showings, and explanatory narration. (Ch. 13) 1961(c).

McTavish, C. L. *Effect of repetitive film showings on learning.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-12, November 1949.

Melton, A. W. The science of learning and the technology of educational methods. *Harvard educ. Rev.*, 1959, *29*, 96-106.

LG Meyer, Susan R. Report on the initial test of a junior high-school vocabulary program. 1960. In A. A. Lumsdaine and R. Glaser (Eds.), *Teaching machines and programmed learning: a source book.* Washington, D. C.: National Education Association, 1960. Pp. 229-246.

Michael, D. N., & Maccoby, N. Factors influencing verbal learning from films under varying conditions of audience participation. *J. exp. Psychol.*, 1953, *46*, 411-418.

L Michael, D. N., & Maccoby, N. Factors influencing the effects of student participation on verbal learning from films: motivating versus practice effects, "feedback," and overt versus covert responding. (Ch. 18) 1961.

Miller, J., & Levine, S. A study of the effects of different types of review and of "structuring" sub-titles on the amount learned from a training film. Washington, D. C.: U. S. Air Force Human Resources Research Laboratory, HRRL Memo. Rept. No. 17, March 1952.

Miller, N. E., et al. Graphic communication and the crisis in education. Audio-Visual Com. Rev., 1957, 5 (3), Special Issue.

Mosteller, F., & Bush, R. R. Selected quantitative techniques. In I. G. Lindzey (Ed.), Handbook of social psychology. Cambridge, Mass.: Addison-Wesley, 1954. Pp. 289-334.

Murnin, J. A. Comparison of training media; transfer of principles involved in a manipulative skill; operation of the aircraft load adjuster slide rule. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-103, September 1955.

Murnin, J. A., VanderMeer, A. W., & Vris, T. Comparison of training media: trainee manipulation and observation of functioning electrical systems versus trainee drawing of schematic electrical systems. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-101, June 1954.

National Workshop. The national program in the use of television in the public schools. New York: The Fund for the Advancement of Education (A report on the National Workshop at Chapel Hill, N. C., June 15-26, 1959), 1959.

Nelson, H. E., & Moll, K. R. Comparison of the audio and video elements of instructional films. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-18, November 1950.

Nelson, H. E., & VanderMeer, A. W. The relative effectiveness of differing commentaries in an animated film on elementary meteorology. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-43, June 1955.

Neu, D. M. The effect of attention-gaining devices on film-mediated learning. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Progress Report No. 14-15-16, 1950.

Newman, S.E. Student vs. instructor design of study method. J. educ. Psychol., 1957, 48, 328-333.

Newman, S. E., & Highland, R. W. The effectiveness of four instructional methods at different stages of a course. Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, Tech. Rept. AFPTRC-TN-56-88, June 1956.

Norford, C. A. Contributions of film summaries to learning from instructional films. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Progress Report No. 13, 1949.

Northrop, D. S. Effects on learning of the prominence of organizational outlines in instructional films. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-33, 1952.

Oinas, F. J. (Ed.) Language teaching today. Int. J. Amer. Linguistics, 1960, 26 (4), Part II.

Otto, H. J., & Flournoy, Frances. Printed materials. Rev. educ. Res., 1956, 26, 115-124.

Pan, S. The influence of context upon learning and recall. J. exp. Psychol., 1926, 9, 468-491.

Park, J. Vocabulary and comprehension difficulties of sound motion pictures. Sch. Rev., 1945, 53, 154-161.

Peterman, J. N., & Bouscaren, N. A study of introductory and summarizing sequences in training film instruction. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Staff Research Memo. (mimeo.), 1954.

LG Peterson, J. C. The value of guidance in reading for information. Trans. Kan. Acad. Sci., 1931, 34, 291-296. [52-58]

Peterson, V., & Schramm, W. How accurately are different kinds of graphs read? Audio-Visual Com. Rev., 1954, 2, 178-189.

LG Porter, D. A. A critical review of a portion of the literature on teaching devices. Harvard educ. Rev., 1957, 27, 126-147. [114-132]

Porter, D. A. Some effects of year long teaching machine instruction. In E. H. Galanter (Ed.), Automatic teaching: the state of the art. New York: Wiley, 1959. Pp. 85-90.

LG Pressey, S. L. A simple apparatus which gives tests and scores--and teaches. Sch. & Soc., 1926, 23 (586), 373-376. [35-41]

LG Pressey, S. L. A machine for automatic teaching of drill material. Sch. & Soc., 1927, 25 (645), 549-552. [42-46]

LG Pressey, S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self instruction. J. Psychol., 1950, 29, 417-447. [69-88]

Rath, G. J., Anderson, Nancy S., & Brainerd, R. C. The IBM Research Center teaching machine project. In E. H. Galanter (Ed.), Automatic teaching: the state of the art. New York: Wiley, 1959. Pp. 117-130.

Rimland, B. Effectiveness of several methods of repetition of films. (Pennsyl-State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-45, May 1955.

Roshal, S. M. Effects of learner representation in film-mediated perceptual-motor learning. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-5, 1949. (See Roshal, 1961.)

Roshal, S. M. The instructional film. In G. Finch (Ed.), Educational and training media: a symposium. Washington, D. C.: National Academy of Sciences - National Research Council, Publication No. 789, 1960. Pp. 114-121.

L Roshal, S. M. Film-mediated learning with varying representations of the task: viewing angle, portrayal of demonstration, motion, and student participation. (Ch. 11) 1961.

Rothkopf, E. Z. Signal similarity and optimal signal arrangement of lesson and drill orders in alphabetic morse code training. In G. Finch & F. Cameron (Eds.), Symposium on Air Force human engineering, personnel, and training research. Washington, D. C.: National Academy of Sciences - National Research Council, Publication No. 516, 1958. Pp. 188-193.

Rothkopf, E. Z. Automated teaching devices and a comparison of two variations of the method of adjusted learning. Murray Hill, N. J.: Bell Telephone Laboratories (mimeo.), 1960.

Saltz, E., & Newman, S. E. The effect of prior learning of symbols of performance in reasoning. Amer. J. Psychol., March 1960, 73, 91-99.

San Francisco State College. An experimental study of college instruction using broadcast television. Project No. 1. San Francisco: San Francisco State College, 1958(a).

San Francisco State College. An experimental study of college instruction using broadcast television. Project No. 2. San Francisco: San Francisco State College, 1958(b).

Saul, E. V., et al. A review of the literature pertinent to the design and use of effective graphic training aids. Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 494-08-1, 1954.

Sheffield, F. D. The role of meaningfulness of stimulus and response in verbal learning. Unpublished doctoral dissertation, Yale University, 1946.

Sheffield, F. D. Perceptual mediation in the learning of organizable sequences: theory and experiment. Air Force Personnel and Training Research Center, Maintenance Laboratory, Tech. Memo. ML-TM-57-14, September 1957. (See Sheffield, 1961.)

L Sheffield, F. D. Theoretical considerations in the learning of complex sequential tasks from demonstration and practice. (Ch. 2) 1961.

L Sheffield, F. D., & Maccoby, N. Summary and interpretation of research on organizational principles in constructing filmed demonstrations. (Ch. 9) 1961.

L Sheffield, F. D., Margolius, G. J., & Hoehn, A. J. Experiments on perceptual mediation in the learning of organizable sequences. (Ch. 8) 1961.

Shettel, H. H., Angell, D., & Lumsdaine, A. A. Self-instructional programs for SAGE system operators: the development, use, and evaluation of two prototype self-instructional programs covering basic job knowledge for SAGE track monitor and intercept director operator positions. Pittsburgh, Pa.: American Institute for Research, April 1961.

Shettel, H. H., Faison, E. J., Roshal, S. M., & Lumsdaine, A. A. An experimental comparison of "live" and filmed lectures employing mobile training devices. Audio-Visual Comm. Rev., 1956, 4, 216-222.

Silverman, R. E. The comparative effectiveness of animated and static transparencies. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. NAVTRADEVCEN 78-1, April 1958.

LG Skinner, B. F. The science of learning and the art of teaching. Harvard educ. Rev., 1954, 24, 86-97. [99-113.]

LG Skinner, B. F. Teaching machines. Science, 1958, 128, 969-977. [137-158.]

Smith, O. W. Comparison of apparent depth in a photograph viewed from two distances. Percept. mot. Skills, 1958, 8 (2).

Smith, O. W., & Gruber, H. Perception of depth in photographs. Percept. mot. Skills, 1958, 4, 307-313.

Smith, O. W., & Resnick, L. Impressions of movements from static line drawings of human figures. Ithaca, N. Y.: Cornell University (unpublished), 1953.

Smith, O. W., Smith, Patricia C., & Hubbard, D. Perceived distance as a function of mode of representation. Ithaca, N. Y.: Cornell University (Unpublished), 1953.

Smith, Patricia C., & Smith, O. W. Ball throwing responses to photographically portrayed targets. J. exp. Psychol., 1961, 62, 223-233.

Smith, P. N., Aukes, L. E., & Lumsdaine, A. A. (Eds.) Conference report. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Tech. Memo. TARL-TM-54-1, 1954.

Stein, J. J. The effect of a pre-film test on learning from an educational sound motion picture. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-35, 1952.

LG Stephens, A. L. Certain special factors involved in the law of effect. Columbus, Ohio: Ohio State University Press, Abstracts of Doctoral Dissertations, 1953, No. 64, 505-511. [89-93]

Stolurow, L. M., & Lumsdaine, A. A. Training characteristics for illumination of animated classroom trainers. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Tech. Memo. TARL-TM-56-1, April 1956.

Stolurow, L. M., & Newman, R. J. A factorial analysis of objective features of printed language presumably related to reading difficulty. Amer. Psychologist, 1956, 11, 403. (Abstract.)

Sulzer, R. L., Lumsdaine, A. A., & Kopstein, F. F. The value of using multiple examples in training film instruction. Washington, D. C.: U. S. Air Force Human Resources Research Laboratory, HRRL Rept. No. 25, May 1952. (See Lumsdaine, Sulzer, & Kopstein, 1961.)

Swanson, R. A. *The relative effectiveness of training aids designed for use in mobile training detachments.* Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, Res. Rept. AFPTRC-TN-54-1, March 1954.

Swanson, R. A., & Aukes, L. E. *Evaluation of training devices for B-47 fuel, hydraulic and rudder power control systems.* Lackland Air Force Base, Texas: Air Force Personnel and Training Research Center, Res. Rept. AFPTRC-TN-56-2, January 1956.

Swanson, R. A., Lumsdaine, A. A., & Aukes, L. E. Two studies in evaluation of maintenance training devices. In G. Finch and F. Camerson (Eds.), *Symposium on Air Force human engineering, personnel, and training research.* Washington, D. C.: National Academy of Sciences - National Research Council, 1956. Pp. 267-275.

Thomson, Louise. The role of verbalization in learning from demonstrations. Unpublished thesis, Yale University, 1944. (Reviewed in *J. educ. Psychol.*, 1946, *37*, 1-12, by M. A. May.)

Tinker, M. A., & Paterson, D. G. *How to make type readable.* New York: Harper, 1940.

Torkelson, G. M. *The comparative effectiveness of a mockup, cutaway, and projected charts in teaching nomenclature and function of the 40 mm. antiaircraft weapon and the Mark 13 type torpedo.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-100, March 1954.

Travers, R. M. W. A study of the relationship of psychological research to educational practice. In R. Glaser (Ed.), *Training research and education.* Pittsburgh, Pa.: University of Pittsburgh Press, 1962.

Twyford, L. *Film profiles.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-23, 1951.

VanderMeer, A. W. *Relative effectiveness of color and black and white in instructional films.* (Pennsylvania State University Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-28, 1952.

VanderMeer, A. W. *Training film evaluation: comparison between two films on personal hygiene: TF8-155 and TF8-1665.* (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-50, October 1953.

Vernon, M. D. The visual presentation of factual data. *Brit. J. educ. Psychol.*, 1950, *20*, 174-185.

Vernon, M. D. *A further study of visual perception.* London: Cambridge University Press, 1952.

Wade, W. L., & Bretz, R. *A report on the experimental use of TV in the UCLA School of Education during 1959-60.* Los Angeles: University of California, University Extension, Educational TV, 1960.

Wald, A. *Statistical decision functions.* New York: Wiley, 1950.

Weiss, J. K. Teaching by television in Hagerstown. *The School Review.* Chicago: University of Chicago Press, 1957. Pp. 466-473.

Weiss, W., & Fine, B. J. Stimulus familiarization as a factor in ideational learning. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-48 (mimeo.), 1955.

L Weiss, W., Sheffield, F. D., & Maccoby, N. Combining practice with demonstration in teaching complex sequences: serial learning of a geometric-construction task. (Ch. 4) 1961.

Witasek, S. Über lesen und rezitieren in ihren beziehungen zum gedachtnis. *Z. Psychol.*, 1907, *44*, 161-185.

Woodworth, R. S. *Experimental psychology*. New York: Holt, 1938.

Wulff, J. J. The teaching effectiveness of a filmed mechanical assembly demonstration with supplementary nomenclature training. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Lab. Note TARL-LN-55-8 (mimeo.), 1955.

L Wulff, J. J., & Emeson, D. L. The relationship between "what is learned" and "how it's taught." (Ch. 30) 1961.

L Wulff, J. J., & Kraeling, Doris G. Familiarization procedures used as adjuncts to assembly-task training with a demonstration film. (Ch. 10) 1961.

Wulff, J. J., Sheffield, F. W., & Kraeling, Doris G. "Familiarization" procedures used as adjuncts to assembly task training with a demonstration film. Air Force Personnel and Training Research Center, Training Aids Research Laboratory, Unpublished Staff Research Memo. (mimeo.), 1954. (See Wulff & Kraeling, 1961.)

Wulff, J. J., & Stolurow, L. M. The role of class descriptive cues in paired-associates learning. *J. exp. Psychol.*, 1957, *53*, 199-206. (See also 1961.)

L Wulff, J. J., & Stolurow, L. M. The role of class-descriptive cues in paired-associate learning. (Ch. 29) 1961.

Zeaman, D. Comments on the automation of teaching. Paper read at Symposium on the Automation of Teaching at the American Psychological Association Annual Convention, Washington, D. C., August 1958.

Zuckerman, J. V. Commentary variations: level of verbalization, personal reference, and phase relations in instructional films on perceptual-motor tasks. (Pennsylvania State University Instructional Film Research Program) Port Washington, N. Y.: U. S. Naval Training Device Center, Office of Naval Research, Tech. Rept. No. SDC 269-7-4, December 1949.

Zuckerman, J. V. Predicting film learning by pre-release testing. Audio-Visual Com. Rev., 1954, 2 (1), 49-56.